ACERCA DA REVERBERAÇÃO EM SINAIS DE VOZ: QUANTIFICAÇÃO PERCEPTUAL E APERFEIÇOAMENTO DE ALGORITMOS DE DESREVERBERAÇÃO

Thiago de Moura Prego

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientadores: Sergio Lima Netto
Amaro Azevedo de Lima

Rio de Janeiro
Maio de 2012

ACERCA DA REVERBERAÇÃO EM SINAIS DE VOZ: QUANTIFICAÇÃO PERCEPTUAL E APERFEIÇOAMENTO DE ALGORITMOS DE DESREVERBERAÇÃO

Thiago de Moura Prego

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Sergio Lima Netto, PhD

Prof. Amaro Azevedo de Lima, PhD

Prof. Eduardo Antonio Barros da Silva, PhD

Prof. Luiz Wagner Pereira Biscainho, DSc

Prof. Rui Seara, DSc

Prof. Tadeu Nagashima Ferreira, DSc

RIO DE JANEIRO, RJ – BRASIL
MAIO DE 2012

Prego, Thiago de Moura

Acerca da Reverberação em Sinais de Voz: Quantificação Perceptual e Aperfeiçoamento de Algoritmos de Desreverberação/Thiago de Moura Prego. – Rio de Janeiro: UFRJ/COPPE, 2012.

XVII, 89 p.: il.; 29, 7cm.

Orientadores: Sergio Lima Netto

Amaro Azevedo de Lima

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2012.

Referências Bibliográficas: p. 75 – 81.

1. Reverberação. 2. Desreverberação. 3. Avaliação de qualidade. I. Netto, Sergio Lima *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

A Deus pelas bênçãos de cada dia
e à minha família pelo suporte
dado em todos estes anos.

Agradecimentos

Meus sinceros agradecimentos:

- aos professores Sergio Lima Netto e Amaro Azevedo de Lima, pela orientação dada durante todo o período do doutorado, pelas oportunidades de participar de projetos bastante interessantes e pela paciência e sabedoria em momentos importantes;
- aos professores Eduardo Antonio Barros da Silva, Luiz Wagner Pereira Biscainho, Rui Seara e Tadeu Nagashima Ferreira, por aceitarem o convite de participação na banca examinadora deste trabalho;
- à HP Labs, pelas inúmeras revisões nos artigos, tornando-os mais bem escritos e com resultados mais sólidos;
- a todas as pessoas que me ajudaram neste projeto por meio de dicas, orientação ou material de estudo.

Thiago de Moura Prego

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

ACERCA DA REVERBERAÇÃO EM SINAIS DE VOZ: QUANTIFICAÇÃO PERCEPTUAL E APERFEIÇOAMENTO DE ALGORITMOS DE DESREVERBERAÇÃO

Thiago de Moura Prego

Maio/2012

Orientadores: Sergio Lima Netto
Amaro Azevedo de Lima

Programa: Engenharia Elétrica

Esta tese tem o objetivo de desenvolver dois avaliadores de qualidade para sinais de fala reverberantes, sendo um com referência (métrica Q_{MOS}) e o outro sem referência (métrica Q_b^{MOS}), e descrever uma metodologia para o ajuste dos parâmetros de algoritmos de desreverberação de sinais de fala.

Utiliza-se uma modelagem do efeito de reverberação do ponto de vista de processamento de sinais, descrita de maneira detalhada. Descreve-se também a base de dados MARDY, utilizada neste trabalho para a validação dos métodos propostos.

É desenvolvida uma base de dados (chamada base NBP) contendo 204 sinais de fala reverberantes, de acordo com três abordagens distintas, possibilitando o desenvolvimento dos métodos para diversos tipos de cenário.

A métrica Q_{MOS} obteve um coeficiente de correlação de 91% e 95% em relação à nota MOS dos sinais das bases NBP e MARDY, respectivamente. A métrica Q_b^{MOS} obteve um coeficiente de correlação de 87% e 91% em relação à nota MOS dos sinais das bases NBP e MARDY, respectivamente. Os resultados obtidos para os avaliadores de qualidade propostos indicam que eles superam outros métodos disponíveis na literatura.

A metodologia para o ajuste de parâmetros de algoritmos de desreverberação de sinais de fala é descrita utilizando-se um algoritmo de dois estágios que melhorou, ao término do processo, em 3% a qualidade estimada através da métrica Q_{MOS} com uma redução de aproximadamente 99% da complexidade computacional.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

ON THE REVERBERATION IN SPEECH SIGNALS: PERCEPTUAL ASSESSMENT AND IMPROVEMENT OF DEREVERBERATION ALGORITHMS

Thiago de Moura Prego

May/2012

Advisors: Sergio Lima Netto
Amaro Azevedo de Lima

Department: Electrical Engineering

The main objective of this thesis is to develop two methods for quality assessment of reverberant speech signals, one with full-reference model (Q_{MOS} measure) and the other with no-reference model (Q_b^{MOS}); and to develop a method to fine tune parameters of speech dereverberation algorithms.

A signal processing based model for reverberation as well as the MARDY database, used for validation of the proposed methods, are described.

A database (called NBP database) containing 204 reverberant speech signals has been developed according to three different approaches, which allows the development of methods for different types of scenario.

The Q_{MOS} measure obtained a correlation coefficient of 91% and 95% regarding the subjective scores for the MARDY and NBP databases, respectively. The Q_b^{MOS} obtained a correlation coefficient of 87% and 91% regarding the subjective scores for the MARDY and NBP databases, respectively. The obtained results indicate that the proposed methods outperform other methods available in the associated literature.

The proposed methodology for fine-tuning of the parameters for speech dereverberation algorithms is described using a two stage algorithm, which has improved, at the end of the process, by 3% the quality estimated by the Q_{MOS} measure with a computational burden reduction of approximately 99%.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

E_{dr}	razão de energia direta sobre reverberante, p. 9
F_s	frequência de amostragem, p. 13
Q	métrica proposta para avaliação de qualidade percebida de sinais reverberantes, p. 17
Q_b	métrica proposta para avaliação cega de qualidade percebida de sinais reverberantes, p. 48
Q_b^{MOS}	métrica Q_b mapeada para a escala MOS, p. 49
Q_{MOS}	métrica Q mapeada para a escala MOS, p. 20
R_{DT}	*reverberation decay tail*, p. 11
T_{60}	período de tempo necessário para a pressão sonora decair 60 dB após o sinal de excitação ter sido interrompido, p. 5
$\bar{J}_d(i)$	variação no tempo da curtose média, p. 61
$\hat{T}_{60}^{(R)}$	Estimativa de Ratnam para T_{60}, p. 41
$\hat{s}(n)$	sinal de fala desreverberado, p. 58
σ_r^2	variância espectral do ambiente, p. 8
$\tilde{s}(n)$	sinal de fala filtrado inversamente, p. 55
$\tilde{s}_p(n)$	resíduo de predição linear do sinal de fala filtrado inversamente, p. 55
$s_r(t)$	sinal de fala reverberante, p. 5
$s_{rp}(n)$	resíduo de predição linear do sinal de fala reverberante, p. 55

Lista de Abreviaturas

ACR	*absolute category rating*, p. 4
DFT	*discrete Fourier transform*, p. 18
EDC	*energy decay curve*, p. 7
FDR	*free-decay region*, p. 29
FFT	*fast Fourier transform*, p. 18
FIR	*finite impulse response*, p. 18
IDFT	*inverse* DFT, p. 18
ITU-T	*International Telecommunication Union - Telecommunication Standardization Sector*, p. 10
LMS	*least mean square*, p. 55
MOS	*mean opinion score*, p. 4
MSE	*mean-square error*, p. 7
NBP	*new Brazilian Portuguese database*, p. 14
ORSMR	*overall reverberation-to-speech modulation energy ratio*, p. 46
RIR	*room impulse response*, p. 5
SEDC	*sub-band* EDC, p. 33
SRMR	*speech-to-reverberation modulation energy ratio, p. 11*
STFT	*short time fourier transform, p. 43*
WPESQ	*wideband perceptual evaluation of speech quality, p. 11*

Capítulo 1

Introdução

Cada vez mais, vem crescendo a demanda por maneiras diferentes de comunicação entre pessoas, independente da distância física entre elas. Suprir essa necessidade é um dos grandes objetivos da área de telecomunicações, na qual pesquisadores tanto da iniciativa privada quanto da pública buscam desenvolver novos mecanismos e dispositivos ou mesmo aperfeiçoar os já existentes. A evolução de tais dispositivos é cada vez mais rápida, o que estimula ainda mais o investimento na área de telecomunicações, realimentando essa evolução e assim sucessivamente.

Áreas de telecomunicações que estão atualmente com grande fomento na pesquisa são:

- Dispositivos móveis para conversações em duas vias com sinais de banda larga (acima da banda telefônica de largura 4 kHz).
- Reconhecimento de sinais de fala.
- Reconhecimento de locutor.
- Dispositivos para apoio a deficientes auditivos.
- Sistemas de teleconferência e telepresença.

Os sistemas referentes a cada uma das áreas citadas são afetados por degradações de diversas naturezas decorrentes do ambiente no qual estão inseridos. Um dos tipos de degradação que afeta tanto a inteligibilidade de sinais de fala quanto o desempenho desses sistemas é a reverberação. Esse efeito está associado às características acústicas de um dado ambiente e em certos níveis torna impraticável o correto funcionamento de equipamentos das áreas citadas anteriormente.

1.1 Proposta de trabalho

Este trabalho tem como principais objetivos o estudo e desenvolvimento de técnicas com e sem referência para a avaliação objetiva da qualidade de sinais de fala sob efeito de reverberação, além de propor aprimoramentos das técnicas para a redução da reverberação (desreverberação) presente nestes sinais.

Para alcançar tais objetivos foi concebida uma base de dados contendo 4 sinais de fala anecoicos e 200 sinais de fala reverberantes originados a partir dos sinais anecoicos. Essa base é composta por três abordagens distintas: artificial, natural e real. O tempo de reverberação T_{60} dos sinais reverberantes dessa base de dados varia entre 120 ms e 920 ms.

Duas técnicas para a avaliação de qualidade são propostas, uma das quais é com referência, isto é, necessita do sinal não degradado além do sinal reverberante e tem como base a técnica proposta por Allen [1, 2], podendo ser utilizada no desenvolvimento e na calibração de sistemas nas áreas citadas. A outra técnica proposta é do tipo sem referência, isto é, necessita apenas do sinal reverberante, com possível aplicação no monitoramento em tempo real destes sistemas.

A metodologia proposta para o aprimoramento das técnicas de desreverberação possui o objetivo de melhorar a qualidade percebida do sinais por elas processados, além de reduzir sua complexidade computacional.

1.2 Organização da tese

O Capítulo 2 define o que é reverberação, descrevendo os principais parâmetros utilizados para a avaliação de qualidade de sinais de fala reverberantes. Esse capítulo também contém a descrição de métodos subjetivos e objetivos de avaliação de qualidade para sinais de fala sob efeito de diversos tipos de degradação. Por fim, a base de dados MARDY, composta por 32 sinais reverberantes e utilizada por vários algoritmos de avaliação de qualidade do tipo em questão, é brevemente descrita.

O Capítulo 3 descreve a base de dados NBP (*new Brazilian Portuguese database*), composta por 204 sinais reverberantes, desenvolvida pelo autor deste trabalho em conjunto com o Laboratório de Processamento de Sinais da COPPE/UFRJ. Também é descrito nesse capítulo o sistema proposto QAreverb, que contém uma técnica com referência para a avaliação de qualidade de sinais reverberantes que têm como base a métrica de Allen. O processo de treinamento do sistema QAreverb e suas peculiaridades práticas são detalhados, assim como a validação do sistema com as bases NBP e MARDY, que mostra o desempenho superior do avaliador de qualidade proposto em relação aos métodos objetivos descritos no Capítulo 2.

O Capítulo 4 propõe um método sem referência para a estimação do tempo de

reverberação de sinais de fala reverberantes, descrevendo seu processo de treinamento. A avaliação do método mostra que seu desempenho é superior a métodos considerados o estado da arte na literatura relacionada.

O Capítulo 5 propõe um método sem referência e não intrusivo para a avaliação de qualidade de sinais de fala reverberantes, com base no método com referência do Capítulo 3. Este método se mostrou superior a todos os outros métodos sem referência para estimação da qualidade percebida de sinais de fala reverberantes disponíveis na literatura e abordados neste trabalho.

O Capítulo 6 descreve uma metodologia para o aprimoramento da qualidade percebida e da complexidade computacional de técnicas de desreverberação para sinais de fala. Esta metodologia é exemplificada a partir de dois conjuntos de propostas de modificações no algoritmo de dois estágios para desreverberação de sinais de fala proposto por Wu e Wang [3]. O primeiro conjunto de alterações é referente ao estágio de filtragem inversa, que almeja remover o efeito das primeiras reflexões, enquanto o segundo conjunto é referente ao estágio de subtração espectral, cujo objetivo é remover os efeitos da reverberação tardia. É feita a combinação dos dois conjuntos de propostas de modificações, além de serem estudadas outras métricas a serem utilizadas como função objetivo da etapa de filtragem inversa.

Por fim, o Capítulo 7 resume o estágio atual da pesquisa, juntamente com suas conclusões, além de apresentar também os próximos passos.

Capítulo 2

Reverberação

2.1 Introdução

Os efeitos da reverberação fazem parte do dia-a-dia de muitas pessoas, sem que a maioria sequer se dê conta disto. Seja em pequenos ambientes como banheiros ou grandes ambientes como o Theatro Municipal do Rio de Janeiro, os efeitos da reverberação estão presentes, ora de maneira perceptualmente desagradável, ora perceptualmente agradável.

É muito comum que engenheiros utilizem o tempo de reverberação como figura de mérito para determinar a quantidade de reverberação presente em um sinal. Encontram-se na literatura dois outros parâmetros que, junto com o tempo de reverberação, são bastante importantes para a descrição do efeito da reverberação. São eles a variância espectral do ambiente e a razão de energia direta sobre reverberante. Estes três parâmetros são definidos com mais detalhes na Seção 2.2.

A Seção 2.3 aborda o tema da avaliação de qualidade de sinais de fala reverberantes. Para tal, é descrito o método de avaliação ACR (*absolute category rating*), que fornece notas subjetivas na escala MOS (*mean opinion score*). Além disso, são descritos os métodos objetivos para avaliação de qualidade perceptual R_{DT} (*reverberation decay tail*), de Allen, recomendação P.563 da ITU-T, SRMR (*speech-to-reverberation modulation energy ratio*) e WPESQ (*wideband perceptual evaluation of speech quality*). Descreve-se também de maneira breve a base de dados MARDY, utilizada na validação de métodos objetivos desta natureza.

Por fim, a Seção 2.4 apresenta as conclusões sobre este capítulo.

2.2 Definição

Reverberação pode ser definida como a modificação de um sinal causada pela resposta acústica do ambiente no qual a fonte de sinal se encontra [4]. Essa resposta

pode ser entendida como sendo as reflexões que o sinal original sofre nas diversas superfícies do ambiente. Um sinal reverberante $s_r(t)$ é constituído, então, pelo sinal original (direto) $s(t)$ e as suas reflexões, que são versões atrasadas e atenuadas do sinal direto. Segundo Mourjopoulos e Hammond [5], o sinal reverberante que chega a um dado sensor é dado por

$$s_r(t) = \int_{-\infty}^{t} h(\tau)s(t-\tau)d\tau, \tag{2.1}$$

em que $h(t)$ é a resposta ao impulso do sistema entre os pontos de transmissão e recepção do sinal, também conhecida como resposta ao impulso do ambiente (RIR - *room impulse response*).

É importante ressaltar que o sinal reverberante é percebido como um único sinal, diferentemente de um sinal que contém eco, no qual as cópias atrasadas e atenuadas do sinal original são percebidas separadamente.

Usualmente divide-se a RIR em duas partes, como mostram as Figs. 2.1 e 2.2:

- Primeiras reflexões: Composta de vários impulsos com amplitudes seguindo, tipicamente, um decaimento exponencial, contendo a maior parte da energia da RIR. Nesse contexto, o primeiro impulso é referente ao componente do som direto.

- Reverberação tardia: Composta pelo restante da RIR, apresenta uma natureza difusa.

Há diversas métricas associadas ao efeito de reverberação como mostram [6–10]. Segundo [9], três dessas métricas demonstraram ser mais importantes para a qualidade percebida por um grupo de ouvintes e, por esta razão, serão detalhadas a seguir. As métricas são o tempo de reverberação, a variância espectral do ambiente e a razão de energia direta sobre reverberante, todas obtidas diretamente da RIR $h(t)$, que pode ser estimada de maneira direta [11], por exemplo através dos algoritmos MLS [12] (*maximum length sequence*), IRS [13] (*inverse repeated sequence*), *time-stretched pulses* [14] ou *sine sweep* [15], ou de maneira indireta, por exemplo, através de métodos de desconvolução.

2.2.1 Tempo de reverberação

O tempo de reverberação de um ambiente especifica a duração para a qual a percepção de um som persiste após ter sido cessado abruptamente [16]. Historicamente, o tempo de reverberação tem sido referenciado como T_{60}, que significa o período de

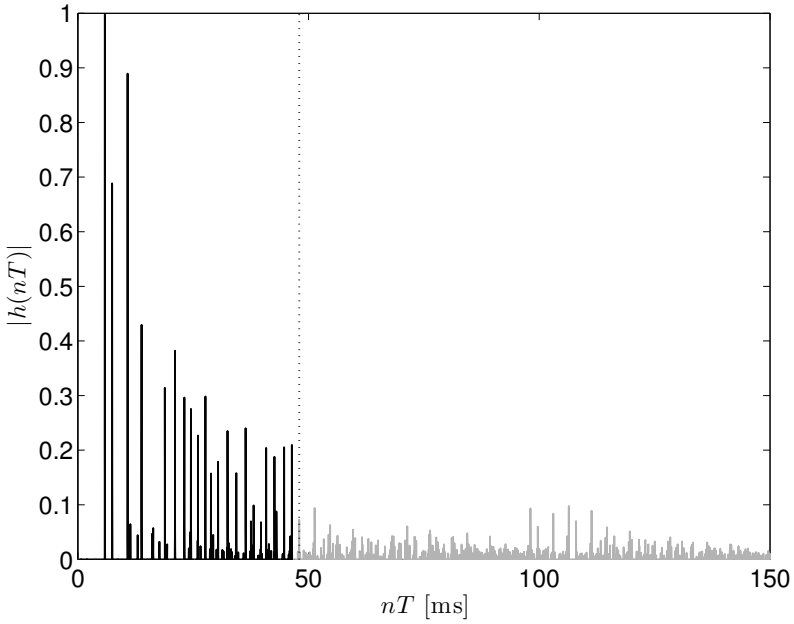


Figura 2.1: RIR artificial no tempo discreto ilustrando a divisão entre primeiras reflexões e reverberação tardia.

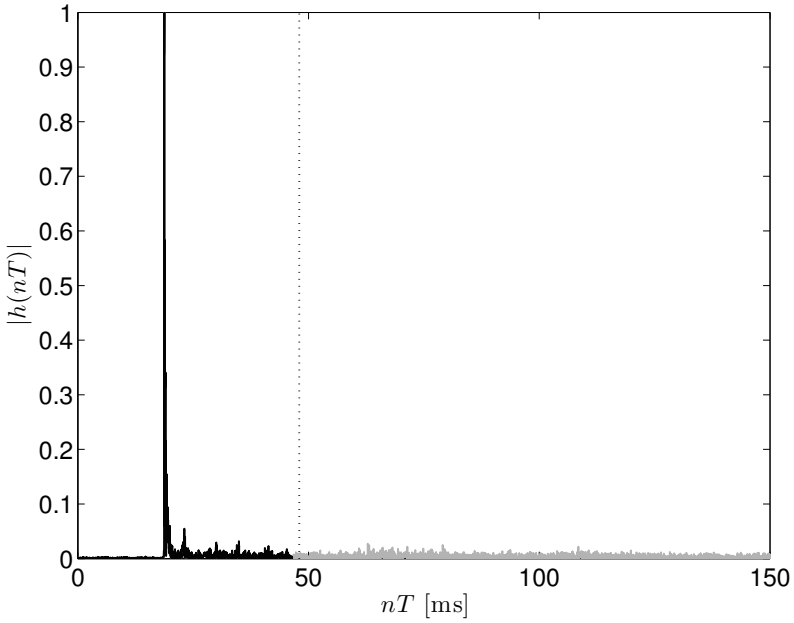


Figura 2.2: RIR real no tempo discreto ilustrando a divisão entre primeiras reflexões e reverberação tardia.

tempo necessário para o som decair 60 dB após o sinal de excitação ter sido interrompido [16]. Na prática, quanto maior o valor de T_{60} mais duradouro é o efeito da reverberação em um ambiente.

Existem inúmeras propostas para se estimar o tempo de reverberação, tais como:

1. Sabine [17] propôs uma fórmula experimental que utiliza somente características geométricas do ambiente e a absorção de suas superfícies.

2. Schroeder [18] propôs um método no qual um pulso breve de banda estreita ou larga é irradiado em um ambiente e a resposta ao impulso para a banda em questão é estimada, a partir da qual se estima o T_{60} através da curva de decaimento da energia (*energy decay curve* - EDC) normalizada, definida mais adiante.

3. Em [19], é proposto o método de ruído interrompido (*interrupted noise method*), no qual uma rajada de ruído de banda estreita ou larga é irradiado no ambiente de teste. Quando o campo sonoro atinge o regime permanente, a fonte de ruído é interrompida e se obtém uma estimativa para o T_{60}. O experimento é realizado diversas vezes e o T_{60} mais provável é determinado utilizando-se uma abordagem de estimação através da máxima verossimilhança.

O método de Schroeder [18] se mostrou superior em relação ao método proposto na ISO 3382 [19], com diversas propostas de modificação como, por exemplo, as referências [20–24].

Assim, dada uma RIR $h(t)$, a EDC é definida como

$$\mathrm{EDC}(t) = 10\log_{10}\left(\frac{\int_t^\infty h^2(\tau)\mathrm{d}\tau}{\int_0^\infty h^2(\tau)\mathrm{d}\tau}\right)\ [\mathrm{dB}], \tag{2.2}$$

em que o denominador garante um valor máximo de 0 dB para a EDC quando $t = 0$. Na escala dB, a EDC pode ser aproximada por uma função de primeira ordem $r(t)$, geralmente iniciando em -5 dB e indo até o ponto de parada no qual se considera que o sinal reverberante atinge o limiar de ruído [18, 22, 23]. Com isto, obtém-se a reta $s(t)$, que é uma versão deslocada de $r(t)$ que passa pelo ponto $(0, 0)$. A partir de $s(t)$ estima-se o valor de T_{60} de modo que $s(T_{60}) = -60$ dB.

A Fig. 2.3 mostra a EDC (linha cheia) para a RIR da Fig. 2.1, na qual a linha tracejada e pontilhada é referente à aproximação de primeira ordem $r(t)$ e a linha tracejada é a aproximação de primeira ordem deslocada $s(t)$. Os pontos $A(0, -5)$ e $B(373, -39)$ foram utilizados para se obter a equação da reta $r(t)$ e o ponto $C \in s(t)$ possui coordenadas $(630, -60)$, logo o $T_{60} = 630$ ms. O ponto B foi obtido através do critério de Lundeby [22], que garante que a aproximação linear escolhida é aquela que produz o menor valor do erro quadrático médio (*mean-square error* - MSE) entre os dados reais e a curva aproximada.

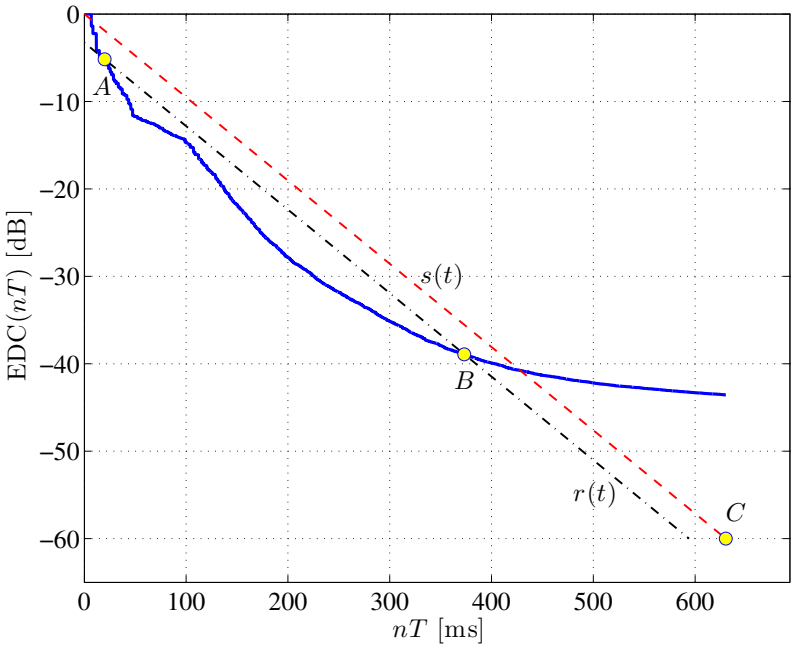


Figura 2.3: Estimação do tempo de reverberação. A curva EDC (linha cheia) gera as aproximação de primeira ordem deslocada $s(t)$, que é utilizada para determinar o ponto C de coordenadas $(T_{60}, -60\text{ dB})$.

2.2.2 Variância espectral do ambiente

Seja $H(f)$ a transformada de Fourier de $h(t)$. Segundo Jetz [25], o nível de intensidade acústica relativa é definido por

$$I(f) = 10\log_{10}\left[\frac{|H(f)|^2}{\int_{-\infty}^{\infty}|H(f)|^2\mathrm{d}f}\right]\ [\text{dB}]. \tag{2.3}$$

A variância espectral do ambiente (*room spectral variance*) σ_r^2 é determinada pela variância de $I(f)$ em dB, isto é,

$$\sigma_r^2 = \int_{-\infty}^{\infty}[I(f) - \overline{I(f)}]^2\mathrm{d}f, \tag{2.4}$$

em que $\overline{I(f)} = \int_{-\infty}^{\infty} I(f)\mathrm{d}f$.

A variância espectral do ambiente caracteriza o efeito de reverberação no domínio da frequência, enquanto o tempo de reverberação o faz no domínio do tempo.

2.2.3 Razão de energia direta sobre reverberante

Seja t_d o instante de tempo associado ao valor máximo de uma RIR $h(t)$. A razão de energia direta sobre reverberante E_{dr} é definida como a razão entre a energia direta E_d (dentro de um pequeno intervalo ao redor de t_d) e a energia reverberante E_r (o restante) de $h(t)$, isto é [26, 27],

$$E_{dr} = \frac{E_d}{E_r} = \frac{\int_{t_d - t_1}^{t_d + t_2} h^2(\tau)\mathrm{d}\tau}{\int_{t_d + t_2}^{\infty} h^2(\tau)\mathrm{d}\tau}, \tag{2.5}$$

em que t_1 é da ordem de 1 ms e t_2 é da ordem de 1,5 ms.

Para reduzir a influência do ruído, Kuster [28] sugere que sejam consideradas apenas os componentes do sinal 20 dB acima do nível de ruído em $h(t)$ e que o acúmulo da energia seja interrompido no mesmo ponto de parada utilizado pelo algoritmo para a estimação de T_{60}.

2.3 Avaliação de qualidade de sinais de fala

Existem duas grandes categorias para métodos de avaliação de qualidade de sinais de fala: métodos subjetivos e métodos objetivos. Os métodos subjetivos são aqueles em que se utilizam seres humanos para avaliar um determinado conjunto de sinais, enquanto os métodos objetivos utilizam algoritmos implementados em máquinas.

É de grande importância saber qual a qualidade percebida de sinais de fala que foram originados ou modificados por algum sistema eletrônico. Por esse motivo, os testes subjetivos são muito utilizados. Geralmente um teste subjetivo utiliza um grupo de pessoas para avaliar sinais de fala e, no final do processo, uma nota média é associada para cada sinal. Como dependem de seres humanos, os testes subjetivos costumam demandar muito tempo, quando comparados com os testes objetivos, além de terem um custo bastante elevado, pela necessidade de um ambiente controlado.

As métricas objetivas são aquelas que, a partir de parâmetros de um sinal de fala, atribuem uma nota à sua qualidade. Apesar de não haver necessariamente uma alta correlação entre uma métrica objetiva e a qualidade percebida por um grupo de seres humanos, há um subgrupo intitulado métricas objetivas perceptuais que tentam emular a nota que um grupo de pessoas daria para um determinado sinal de fala. Essas métricas têm a vantagem de não dependerem de pessoas (na sua execução), sendo ordens de grandeza mais rápidas do que os testes subjetivos. Devido ao pe-

queno tempo necessário (em relação aos testes subjetivos) para a execução de testes objetivos, esses costumam ser utilizados tanto na criação quanto no monitoramento de sistemas cujas saídas são sinais de fala.

Há diversos trabalhos detalhados sobre testes subjetivos e objetivos, tais como as referências [29, 30].

2.3.1 MOS - *mean opinion score*

No ano de 1996, a ITU-T (*International Telecommunication Union - Telecommunication Standardization Sector*) padronizou testes subjetivos para a avaliação da qualidade percebida de sinais de fala [31]. Nesses testes, um grupo de ouvintes escutam sinais de fala e dão uma nota para a qualidade percebida. Um dos testes mais utilizados é conhecido como ACR (*absolute category rating*), no qual cada ouvinte escuta sinais compostos por poucas sentenças descorrelacionadas (geralmente duas sentenças com um intervalo de 500 ms entre elas) e lhes atribui uma nota de 1 a 5 conforme a escala MOS (*mean opinion score*) apresentada na Tabela 2.1.

Tabela 2.1: Tabela MOS.

MOS	Qualidade do sinal de fala
5	Excelente
4	Bom
3	Regular
2	Ruim
1	Péssimo

2.3.2 Métrica de Allen

O tempo de reverberação T_{60} é, sem dúvidas, o parâmetro mais importante para se quantificar o efeito da reverberação em um dado sinal de fala. Em [1], Allen propôs uma métrica para estimar a qualidade de um sinal de fala percebida por um grupo de indivíduos, combinando o tempo de reverberação com a variância espectral do ambiente σ_r^2,

$$P = P_{\max} - \sigma_r^2 T_{60}, \tag{2.6}$$

em que $P_{\max}$ é um valor arbitrário que limita o valor máximo de P, lembrando que σ_r^2 e T_{60} são números reais não-negativos. Quanto maior o valor de P, maior é a qualidade de um sinal de fala. Em [2], essa métrica é validada experimentalmente para diversos sinais de fala.

2.3.3 Métrica R_{DT}

Em [32], Wen e Naylor utilizam o modelo de RIR proposto por Habets [33],

$$h(t) = b(t)A\mathrm{e}^{-\lambda t}, \tag{2.7}$$

em que $b(t)$ é ruído branco gaussiano de média zero, A é um fator de escala e λ é a taxa de decaimento associada ao tempo de reverberação.

Wen e Naylor, então, definem a métrica R_{DT} (*reverberation decay tail*), que é obtida ao se calcular diferenças espectrais na escala Bark [34] (*Bark spectral differences*) entre a referência e o sinal reverberante, em cada *bin* espectral na escala Bark, isto é,

$$R_{DT} = \frac{\overline{A}}{\overline{\lambda} \cdot \overline{E_d}}, \tag{2.8}$$

em que $\overline{A}$ é a média da estimativa de A, $\overline{\lambda}$ é a média da taxa de decaimento e $\overline{E_d}$ é a média da energia direta dos *bins* espectrais na escala Bark. Quanto menor for R_{DT}, maior será a qualidade do sinal de fala.

2.3.4 Métrica SRMR (*speech-to-reverberation modulation energy ratio*)

Em [35], Falk et al. investigaram uma representação de modulação espectral para a avaliação sem referência de qualidade e inteligibilidade de sinais de fala reverberantes e desreverberados com largura de banda de 4 kHz ou 8 kHz. Essa representação é obtida através da utilização do banco de filtros Gammatone auditivos [36] e é utilizada para o cálculo da métrica SRMR,

$$\mathrm{SRMR} = \frac{\sum_{k=1}^{4} \bar{\varepsilon}_k}{\sum_{k=5}^{K^*} \bar{\varepsilon}_k}, \tag{2.9}$$

em que $\bar{\varepsilon}_k$ é a energia média da k-ésima banda de modulação e K^* é escolhido para cada sinal. Quanto maior for a métrica SRMR, maior será a qualidade do sinal de fala.

2.3.5 Métrica WPESQ

A Recomendação P.862.2 [37] da ITU-T, conhecida como WPESQ (*wideband perceptual evaluation of speech quality*), descreve uma extensão para a aplicação do algoritmo PESQ [38] (originalmente desenvolvido para sinais de banda telefônica de largura 3,4 kHz) em sinais com largura de banda de até 7 kHz. O algoritmo

WPESQ, assim como o PESQ, foi desenvolvido para estimar a qualidade percebida de sinais contendo as seguintes degradações:

- Codificação de forma de onda.
- Codificação paramétrica ou híbrida a partir de 4 kbps.
- Erros no canal de transmissão.
- Perda de pacotes.

O algoritmo WPESQ é do tipo com referência, isto é, o sinal original (sem degradação) é comparado com o sinal degradado para se obter a nota objetiva final. Essa nota WPESQ é mapeada em uma nota MOS a partir da seguinte função:

$$\text{MOS} = 0{,}999 + \frac{4}{1 + \text{e}^{-1{,}3669\text{WPESQ}+3{,}8224}}. \tag{2.10}$$

2.3.6 Recomendação ITU-T P.563

A Recomendação P.563 [39] da ITU-T descreve um algoritmo sem referência para a estimação da nota MOS que seria obtida por um teste subjetivo do tipo ACR. O algoritmo foi desenvolvido para sinais em banda telefônica (de largura 3,4 kHz) transmitidos por redes telefônicas suscetíveis aos seguintes problemas:

- Características acústicas do ambiente.
- Ruído ambiente no transmissor.
- Características acústicas do terminal transmissor.
- Características elétricas e de codificação do terminal transmissor.
- Erros no canal de transmissão.
- Perda de pacote e correção de perda de pacotes de codificadores do tipo CELP (*code excited linear prediction*).
- Mudança de taxa de bits para codificadores com taxa variável.
- Efeitos da variação do atraso em testes do tipo ACR.
- *Time warping* de curto e longo termo em sinais de fala.
- Sistemas de transmissão, tais como canceladores de eco e redutores de ruído.
- Sistemas em que há apenas uma fonte ativa de sinais de fala por vez.

2.3.7 Base de dados MARDY

A base de dados MARDY [40] (*Multichannel Acoustic Reverberation Database at York*) é composta por 16 sinais de fala reverberantes gravados diretamente em um auditório e por uma versão desreverberada de cada um dos 16 sinais utilizando o algoritmo de atraso-e-soma [41], totalizando 32 sinais de fala com frequência de amostragem $F_s = 16$ kHz. Essa base contém 2 locutores, um do sexo masculino e outro do sexo feminino, 4 valores de distância fonte-microfone ($d = 1, 2, 3, 4$ m) e 2 tipos de painéis de parede (reflexivo e absortivo), que correspondem a valores estimados de tempo de reverberação T_{60} de 447 ms e 291 ms, respectivamente.

A base de dados MARDY foi, provavelmente, a primeira desenvolvida com o propósito de se avaliar os efeitos da reverberação em sinais de fala. Entretanto, essa base contém apenas um pequeno número de sinais, todos gravados em uma mesma sala. Tais características motivaram a criação de uma base de dados maior e mais geral, descrita na Seção 3.2.

2.4 Conclusões

Neste capítulo, foram abordados os efeitos da reverberação na percepção da qualidade de sinais de fala. Esses efeitos podem ser estimados com bastante eficácia a partir dos parâmetros tempo de reverberação, variância espectral do ambiente e energia direta sobre reverberante, sendo apresentadas maneiras de obtenção de tais parâmetros. Foram apresentados também um método subjetivo para a avaliação de qualidade percebida que utiliza a escala MOS e os métodos objetivos R_{DT}, de Allen, WPESQ, SRMR e ITU-T P.563, que utilizam alguns dos parâmetros mencionados ou variantes destes parâmetros. Por fim, foi apresentada a base de dados MARDY, composta por 32 sinais reverberantes, utilizada na literatura para a validação de métodos objetivos de avaliação de qualidade de sinais de fala sob efeito de reverberação.

Capítulo 3

Sistema QAreverb

3.1 Introdução

Este capítulo tem por objetivo descrever um sistema completo com referência para a avaliação de qualidade de sinais de fala sob efeitos de reverberação, bem como todo o processo de treinamento deste sistema. Para tal, primeiramente é descrita, na Seção 3.2, a base de dados NBP, composta por 204 sinais de fala sob efeito de três conjuntos de reverberação: artificial, natural e real.

A Seção 3.3 descreve o sistema QAreverb, detalhando as suas etapas que são: pré-processamento, estimação da RIR $h(t)$, estimação dos parâmetros T_{60}, σ_r^2 e E_{dr}, cálculo de uma nota objetiva que utiliza os três parâmetros estimados e mapeamento linear da nota objetiva calculada.

Na Seção 3.4, é detalhado o processo de treinamento do sistema QAreverb e são mostrados resultados experimentais que comprovam a maior eficácia do sistema em relação a outros métodos.

Por fim, a Seção 3.5 apresenta um resumo do que foi discutido, além das conclusões do capítulo.

3.2 Base de dados

Para o desenvolvimento da métrica QAreverb e para a comparação com outros métodos da literatura foi criada a base de dados nova Base para Português Brasileiro (*new Brazilian Portuguese database* - NBP), composta por sinais afetados por reverberação.

A base NBP foi completamente desenvolvida a partir de 4 sinais de fala anecoicos (T_{60} da ordem de 30 ms) gravados digitalmente com frequência de amostragem $F_s = 48$ kHz, dois deles com um locutor masculino e os outros dois com um locutor feminino. Cada sinal anecoico é composto por duas frases curtas em Português

Brasileiro, separadas por aproximadamente 1,7 s de silêncio, com duração média de 8,4 s. O efeito de reverberação foi aplicado a cada um dos 4 sinais anecoicos através de três abordagens distintas:

- Reverberação artificial: Neste método, o efeito de reverberação foi emulado a partir de 6 RIRs geradas artificialmente, dando origem a um total de 24 sinais reverberantes obtidos pela convolução dos 4 sinais anecoicos com as 6 RIRs. Nessas RIRs, as primeiras reflexões foram modeladas através do método das imagens [42], com uma distância fonte-microfone fixa de 1,8 m em uma sala virtual de dimensões comprimento×largura×altura= $4 \times 3 \times 3$ m. Em relação à reverberação tardia, o método de rede realimentada de atrasos (*feedback delay network*) [43] foi utilizado para emular tempos de reverberação $T_{60} = 200, 300, 400$ ms e uma versão modificada do método de Gardner [4, 44] (originalmente utilizado para valores de T_{60} superiores a 400 ms) para emular tempos de reverberação $T_{60} = 500, 600, 700$ ms.

- Reverberação natural: Nesta abordagem, 17 RIRs da base AIR (*Aachen impulse response database*) descrita em [45] foram utilizadas para gerar 68 sinais reverberantes obtidos através da convolução com os 4 sinais anecoicos. Essas RIRs foram obtidas de 4 salas distintas com diferentes distâncias fonte-microfone d, como mostrado na Tabela 3.1. Na coluna dimensões, (C×L×A) indica (comprimento×largura×altura).

Tabela 3.1: Características das salas do efeito de reverberação natural da base NBP

Tipo de sala	Dimensões (C×L×A) [m]	T_{60} [ms]	d [m]
cabine	3,0×1,8×2,2	120	0,5; 1; 1,5
escritório	5,0×6,4×2,9	430	1; 2; 3
reunião	8,0×5,0×3,1	230	1,45; 1,7; 1,9; 2,25; 2,8
aula	10,8×10,9×3,15	780	2,25; 4; 5,6; 7,1; 8,7; 10,2

- Reverberação real: Neste método, os quatro sinais foram reproduzidos em uma caixa de som e gravados por um microfone dentro de 7 salas com diferentes características de reverberação. Em cada sala, 4 valores de distâncias fonte-microfone d foram consideradas, com exceção da menor sala, na qual apenas 3 distâncias foram consideradas, como mostrado na Tabela 3.2. Nessa tabela, os valores de T_{60} são as médias obtidas para as diferentes distâncias fonte-microfone em cada sala. Os altos valores do tempo de reverberação associados às salas reunião2 e escritório2 são oriundas das características altamente reflexivas das paredes dessas salas. Esta abordagem gerou 108 sinais reverberantes. Os desenhos esquemáticos das 7 salas utilizadas se encontram no Anexo A.

Tabela 3.2: Características das salas utilizadas para se obter o efeito de reverberação real da base NBP

Tipo de sala	Dimensões (C×L×A) [m]	T_{60} [ms]	d [m]
cabine	2,1×1,8×2,4	140	0,5; 1; 1,5
escritório1	7,4×5,0×2,7	390	1; 2; 3; 4
aula1	15,0×10,0×4,0	570	1; 2; 3; 4
reunião1	10,0×4,8×3,2	650	1; 2; 3; 4
aula2	16,5×8,2×3,5	700	1; 2; 3; 4
reunião2	9,0×7,3×3,5	890	1; 2; 3; 4
escritório2	7,4×4,8×4,3	920	1; 2; 3; 4

Com o objetivo de se obter um *ground truth* para a qualidade estimada, um teste subjetivo foi feito para os 204 sinais da base NBP (24 com reverberação artificial, 68 com reverberação natural, 108 com reverberação real e 4 anecoicos) com base no teste ACR da ITU-T descrito na Seção 2.3.1. Neste teste, cada sinal foi ouvido e avaliado por 30 diferentes ouvintes que deveriam dar uma nota entre 1 e 5, com 0,1 de passo, como mostra a Fig. 3.1. Como recomendado por [30], os rótulos sugeridos pela norma P.800 da ITU-T foram removidos, pois a sua presença causa uma não linearidade na escala, uma vez que cada avaliador interpreta de maneira singular o significado de cada rótulo. Caso não haja rótulos, cada avaliador tende a usar a escala de maneira linear. A primeira página do formulário utilizado no teste subjetivo se encontra no Anexo B.

Figura 3.1: Escala de notas utilizada no teste subjetivo da base NBP.

Cada ouvinte participou de duas sessões em dias diferentes, cada uma com cerca de 16 min de duração, na qual 112 sinais foram ouvidos e avaliados. Os primeiros 10 sinais não fazem parte da base NBP e foram utilizados para treinar os ouvidos da cada avaliador, uma vez que cobriam toda a faixa de T_{60} presente na base NBP. Os outros 102 sinais foram apresentados em uma ordem aleatória (gravada em um arquivo para o registro das notas).

Para atribuir uma nota MOS para cada sinal, os valores atípicos (*outliers*) foram removidos utilizando o critério de 3 desvios padrão em torno da média da nota de cada sinal. Do total de 6120 notas atribuídas à base completa, apenas 9 foram removidas por serem consideradas valores atípicos, e não houve caso de mais de um valor atípico para um mesmo sinal. A nota MOS junto com o desvio padrão correspondente para cada sinal da NBP está representada na Fig. 3.2, em que os

sinais foram ordenados de forma crescente em relação à nota MOS. As margens de erro resultantes são comparáveis quantitativamente com notas subjetivas de sinais da ITU-T utilizados no teste do algoritmo de avaliação de qualidade de sinais contendo ruído de codificação PESQ [38] (*Perceptual Evaluation of Speech Quality*).

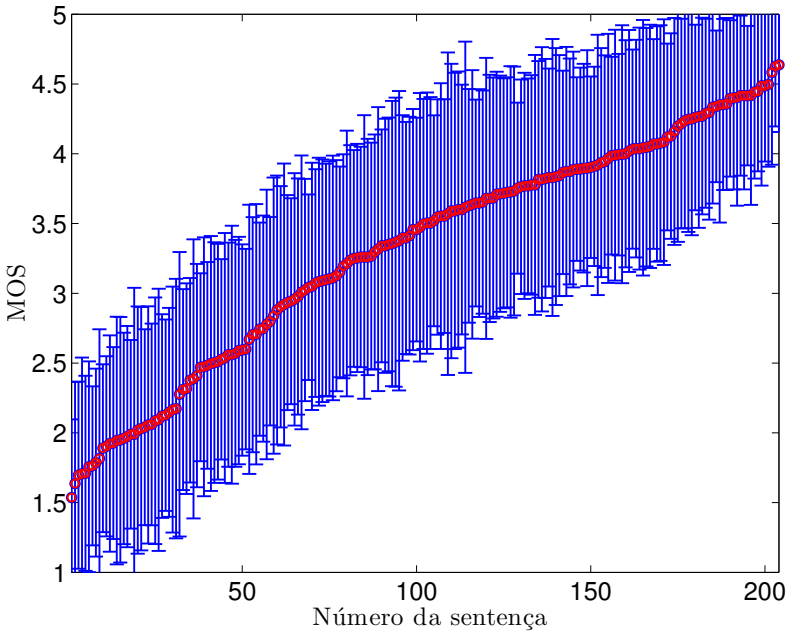


Figura 3.2: Nota subjetiva MOS e desvio padrão correspondente para cada sinal da base NBP.

3.3 QAreverb

Como mostram diversos estudos encontrados na literatura, o tempo de reverberação é o parâmetro mais importante para se quantificar o efeito de reverberação. Em [42] e [1], Allen propôs uma métrica para a quantificação da qualidade percebida de sinais de fala afetados por reverberação, a qual combina o tempo de reverberação T_{60} e a variância espectral do ambiente σ_r^2. Segundo Cole et al. [46] e Griesinger [47], a razão de energia direta sobre reverberante E_{dr} exerce um papel fundamental na avaliação da inteligibilidade de sinais de fala em ambientes fechados, fornecendo ao ouvinte informação acerca da localização e da distância de uma dada fonte sonora.

Experimentos práticos feitos por Lima et al. [4] também indicam a importância da E_{dr} na percepção subjetiva dos efeitos da reverberação. Apesar de σ_r^2 ser bastante relacionada a E_{dr}, quando $E_{dr} \geq 1$, esses dois parâmetros se tornam fortemente dissociados para grandes distâncias fonte-microfone [48], isto é, quando $E_{dr} < 1$ [25].

Com base em tal afirmação, propõe-se uma nova métrica Q para a avaliação de qualidade perceptual para sinais de fala com largura de banda de até 24 kHz afetados por reverberação em que são combinados os parâmetros T_{60}, σ_r^2 e E_{dr}, incorporando

a razão de energia E_{dr} à métrica de Allen da seguinte maneira:

$$Q = -\frac{T_{60}\sigma_r^2}{E_{dr}^{\gamma}}, \tag{3.1}$$

em que γ é um expoente determinado experimentalmente no estágio de treinamento do sistema. O caso especial no qual $\gamma = 0$ corresponde à métrica original de Allen [42] com $P_{\max} = 0$.

O sistema geral para a avaliação de qualidade perceptual de sinais de fala reverberantes que utiliza a métrica Q pode ser implementado no domínio do tempo discreto n, para uma dada frequência de amostragem F_s, como mostra a Fig. 3.3. Este sistema recebe o sinal de fala limpo (sem reverberação) $x_c(n)$ e o sinal de fala reverberante $x_r(n)$, ambos em versões discretizadas de seus correspondentes no domínio do tempo contínuo. Ambos os sinais são utilizados na obtenção de uma estimativa $\hat{h}(n)$ da resposta ao impulso do ambiente, a partir da qual os parâmetros T_{60}, σ_r^2 e E_{dr} são estimados. De posse dessas três estimativas, pode-se determinar a métrica Q segundo (3.1).

No tempo discreto, segundo Lima et al. [4], a RIR geralmente é modelada como uma resposta ao impulso de duração finita (FIR - *finite impulse response*). Reescrevendo (2.1), obtém-se que o sinal de fala reverberante é a convolução linear entre o sinal de fala direto e a resposta ao impulso do ambiente:

$$s_r(n) = \sum_{i=0}^{N} h(i)s(n-i) \tag{3.2}$$

Neste esquema, o papel de cada bloco da Fig. 3.3 é detalhado a seguir:

- PRÉ-PROCESSAMENTO: Remove o nível médio dos sinais $x_c(n)$ e $x_r(n)$, gerando os sinais pré-processados $\tilde{x}_c(n)$ e $\tilde{x}_r(n)$.

- DESCONVOLUÇÃO: Estima a RIR $\hat{h}(n)$ através da desconvolução entre os sinais $\tilde{x}_c(n)$ e $\tilde{x}_r(n)$ segundo

$$\hat{h}(n) = \text{IDFT}\left[\frac{\text{DFT}[\tilde{x}_r(n)]}{\text{DFT}[\tilde{x}_c(n)]}\right], \tag{3.3}$$

  em que os operadores DFT[·] e IDFT[·] representam a transformada discreta de Fourier e a transformada discreta de Fourier inversa, respectivamente. Limitando-se as estimações a 1 segundo, as transformadas podem ser obtidas por algoritmos rápidos, tais como a transformada discreta de Fourier rápida FFT (*fast Fourier transform*) de tamanho dado pela potência de 2 maior do que ou igual a $(F_s + L_c - 1)$, em que L_c é o comprimento (em amostras) do sinal $\tilde{x}_c(n)$.

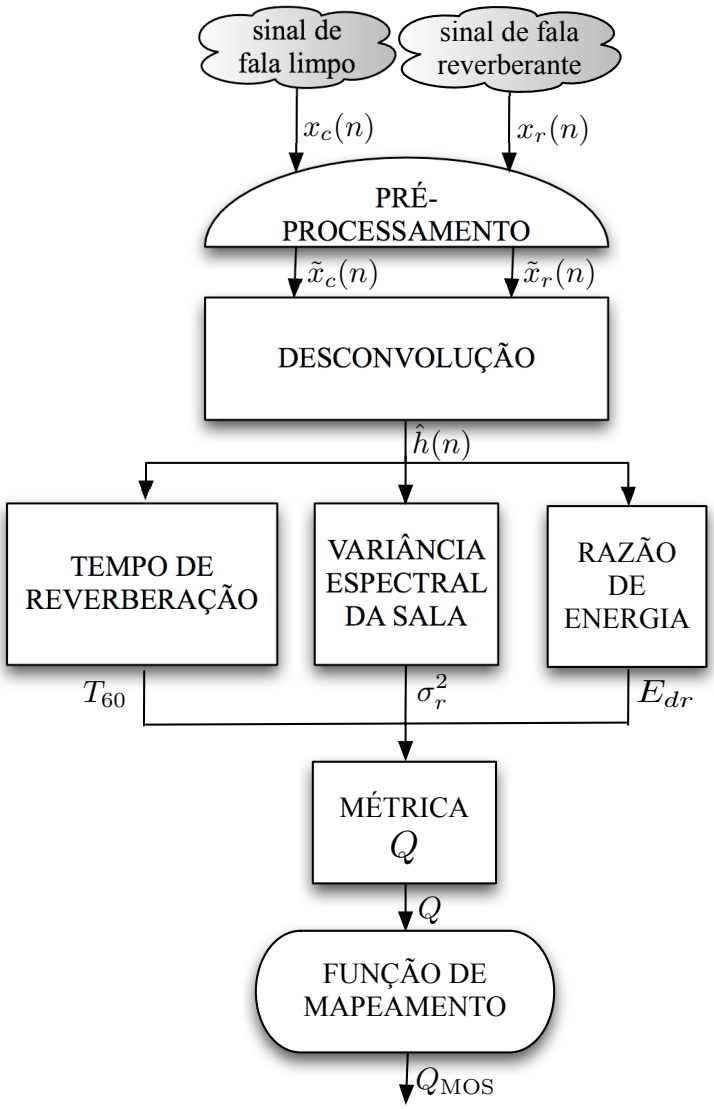


Figura 3.3: Diagrama de blocos do sistema QAreverb, que utiliza a nova métrica Q_{MOS}.

Ao implementar (3.3), os problemas oriundos de pequenos valores do denominador devem ser resolvidos. A abordagem proposta é usar um limiar $\epsilon > 0$ no valor absoluto de $\tilde{X}_c(k) = \mathrm{DFT}[\tilde{x}_c(n)]$ e, caso $|\tilde{X}_c(k)| < \epsilon$, forçar $\tilde{X}_c(k) = \epsilon$, isto é,

$$\text{para todo } k: \text{ se } |\tilde{X}_c(k)| < \epsilon \Rightarrow \tilde{X}_c(k) \equiv \epsilon, \tag{3.4}$$

em que $A \equiv B$ significa que A é substituído por B. A influência desse parâmetro ϵ é discutida mais adiante na Seção 3.4.1.

- TEMPO DE REVERBERAÇÃO: Estima o parâmetro T_{60} a partir de $\hat{h}(n)$, através do algoritmo proposto por Karjalainen et al. [24] (seu código para MATLAB® foi fornecido pelos autores).

- VARIÂNCIA ESPECTRAL DO AMBIENTE: Determina σ_r^2 associada à RIR

estimada $\hat{h}(n)$ através de uma versão de (2.4) no domínio da frequência discreta. Em outras palavras, σ_r^2 é a variância do nível de intensidade acústica relativa $I(k)$, que é dado por

$$I(k) = 10 \log_{10} \left[\frac{|\hat{H}(k)|^2}{\overline{|\hat{H}(k)|^2}} \right] \text{ [dB]}, \tag{3.5}$$

em que $\hat{H}(k) = \text{DFT}[\hat{h}(n)]$.

- RAZÃO DE ENERGIA: Calcula a razão da energia direta sobre reverberante E_{dr} a partir de $\hat{h}(n)$ utilizando a versão no domínio do tempo discreto de (2.5), em que t_d é o instante de tempo (em segundos) associado ao valor máximo de $|\hat{h}(n)|$, isto é,

$$t_d = \frac{\arg \left\{ \max_n |\hat{h}(n)| \right\}}{F_s}. \tag{3.6}$$

- MÉTRICA Q: Determina a métrica Q conforme (3.1).

- FUNÇÃO DE MAPEAMENTO: Mapeia o valor de Q na escala MOS, utilizando um modelo polinomial de terceira ordem semelhante ao utilizado pela Recomendação P.563 da ITU-T [39]

$$\bar{Q} = x_1 Q^3 + x_2 Q^2 + x_3 Q + x_4, \tag{3.7}$$

em que os coeficientes x_1, x_2, x_3 e x_4 são determinados durante o estágio de treinamento do sistema. Na prática, diferentes notas podem ser dadas para um mesmo sinal de fala reverberante caso diferentes intervalos de tempo de reverberação sejam considerados em um dado teste subjetivo. Por esse motivo, este procedimento é seguido por um ajuste linear de escala [30], dado por

$$Q_{\text{MOS}} = \alpha \bar{Q} + \beta, \tag{3.8}$$

em que α e β podem ser determinados por um subconjunto da base de treinamento. O ajuste linear de escala diminui o erro quadrático médio (MSE) entre as métricas objetiva e subjetiva, sem alterar o fator de correlação associado [49].

3.4 Treinamento e validação do sistema QAreverb

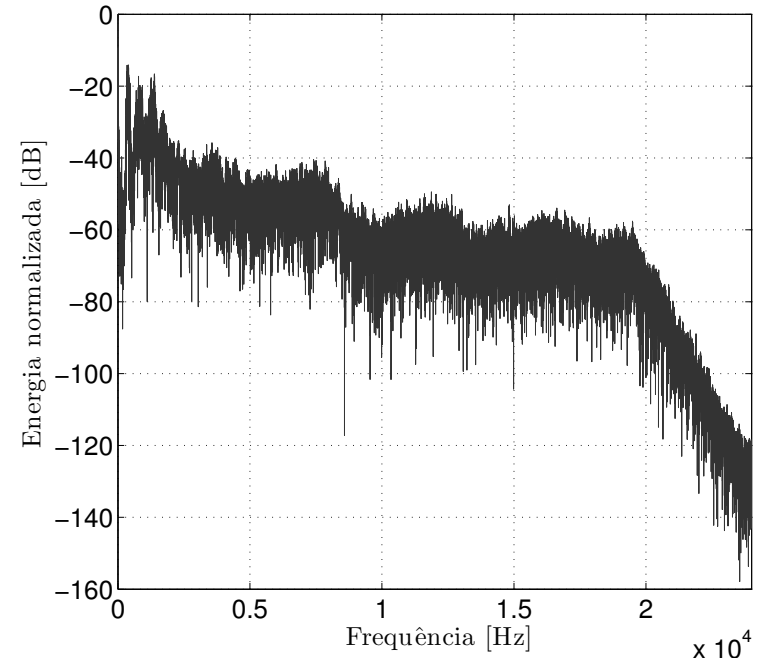
3.4.1 Escolhendo o valor de ϵ

Diferentes configurações de gravação acarretam diferentes formatos espectrais, como exemplificado nas Figs. 3.4(a) e (b), que ilustram os gráficos da DFT de um sinal da base NBP e de um sinal da base MARDY. A energia foi normalizada para que o máximo de cada gráfico seja 0 dB.

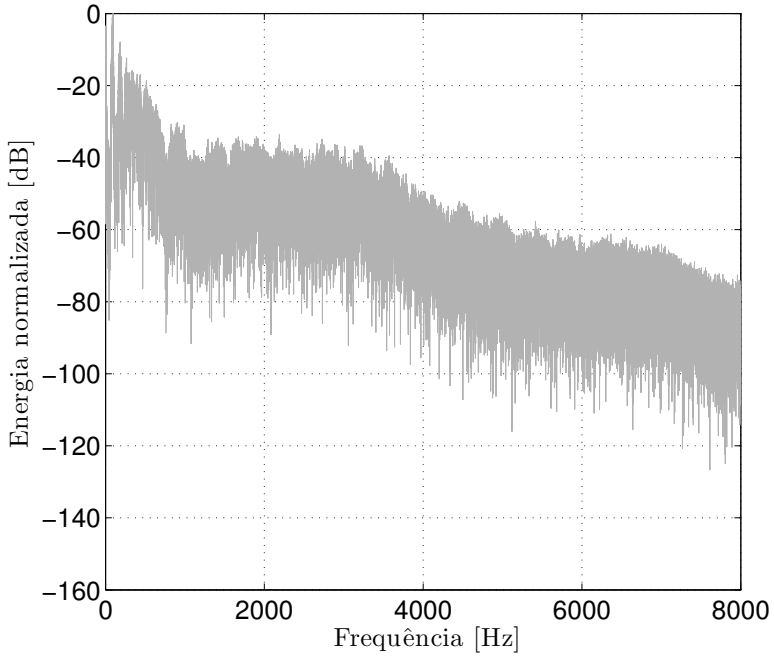
Se considerarmos que a banda de rejeição dos filtros anti-*aliasing* dos conversores A/D, aos quais os sinais das bases MARDY e NBP foram submetidos, começam na região de -60 dB de queda, vemos que isso ocorre por volta de 4 kHz (25% da frequência de amostragem) para a MARDY e por volta de 20 kHz (aproximadamente 42% da frequência de amostragem) para a NBP. Esses gráficos representativos mostram que o filtro anti-*aliasing* utilizado na base de dados MARDY possui uma região transitória relativamente maior do que o filtro utilizado na base NBP. Por causa disso, o valor de ϵ de (3.4) deve ser ajustado para cada base de dados para reduzir os erros numéricos associados ao processo de estimação da RIR, isto é, erros numéricos de (3.3).

Na prática, esse procedimento pode ser feito de maneira bastante simples e robusta ao considerar o comportamento de σ_r^2 como uma função de ϵ, como mostrado na Fig. 3.5 para ambas as bases de dados. Esses gráficos indicam que pequenos valores de ϵ levam a um comportamento plano de σ_r^2 similar para todos os sinais de uma dada base de dados, até um dado limiar a partir do qual a variância espectral do ambiente muda de maneira drástica. Observando (2.3), vemos que a média de $|H(f)|^2$ se encontra no denominador da fração e, por causa da maior quantidade de valores de $|H(f)| = \epsilon$, quanto maior o valor de ϵ, menor será o valor de $\int_{-\infty}^{\infty} |H(f)|^2 \mathrm{d}f$, levando aos padrões crescentes para σ_r^2 mostrados nas Figs. 3.5 (a) e (b). Na Fig. 3.5 (a) há um grupo de sinais com $\sigma_r < 8$ (composto por alguns sinais do perfil artificial com o locutor feminino) para $\epsilon = 10^{-2}$ e outro grupo de sinais com $\sigma_r^2 > 10$ (composto pelos demais sinais não anecoicos). Na Fig. 3.5 (b) há um grupo de sinais com $\sigma_r^2 > 10$ (composto pelos 16 sinais desreverberados) e outro grupo de sinais com $\sigma_r < 9$ (composto pelos 16 sinais reverberantes) para $\epsilon = 10^{-1}$. O diferente valor de ϵ para o joelho dos gráficos de σ_r^2 para as bases NBP e MARDY se dá pela diferença na riqueza espectral das bases.

Além disso, esses gráficos também indicam que todos os sinais não anecoicos de uma dada base apresentam variações (em relação a ϵ) de σ_r^2 similares. Esse padrão, quando determinado para um único sinal de uma dada base de dados, pode ser utilizado para se obter um valor experimental para ϵ a ser utilizado em todos os sinais daquela base de dados.



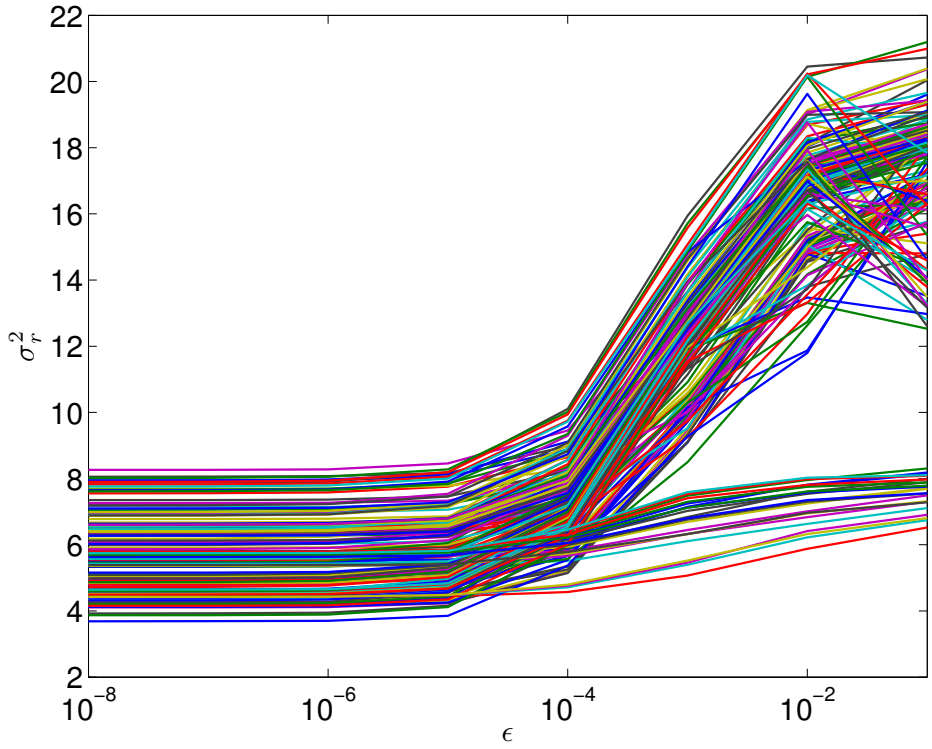
(a)



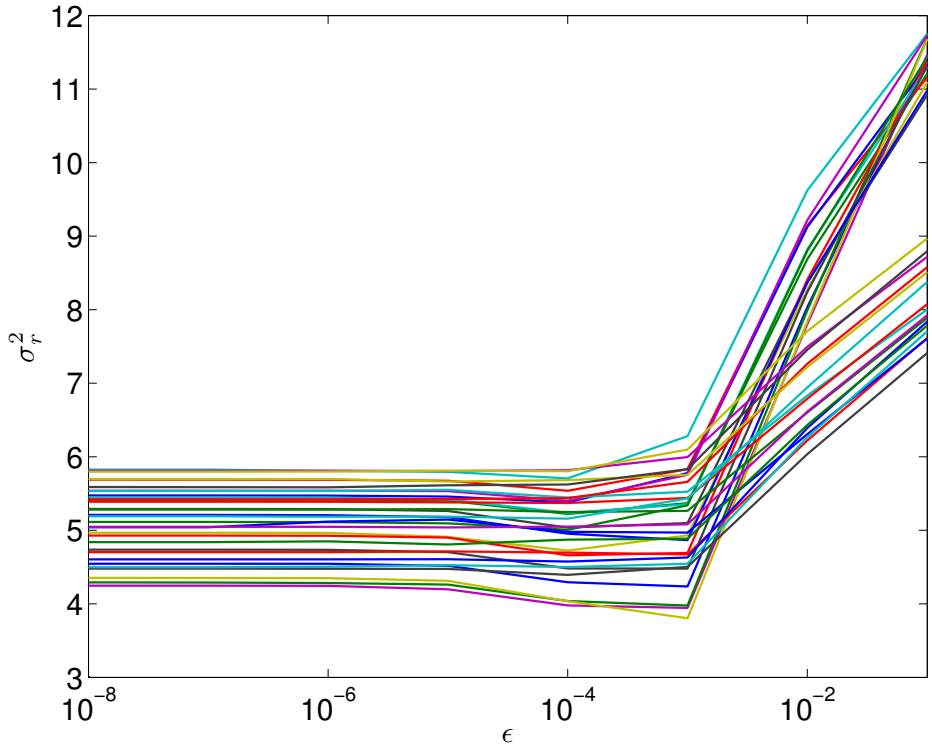
(b)

Figura 3.4: Conteúdo espectral normalizado de um sinal sem reverberação oriundos das bases: a) NBP (cinza escuro) e b) MARDY (cinza claro), como aparecem no denominador de (3.3).

Como pequenos valores de ϵ acarretam erros numéricos em (3.3), afetando a estimativa dos três parâmetros referentes a reverberação, uma estratégia coerente é escolher um valor de ϵ dentro da região plana de σ_r^2 e próximo do limiar de mudança de perfil. Para o caso das bases NBP e MARDY, os valores para o limiar são $\epsilon = 10^{-5}$ e $\epsilon = 10^{-3}$, respectivamente. O valor de ϵ muito maior utilizado para



(a)



(b)

Figura 3.5: Variação de σ_r^2 em relação ao parâmetro regulador ϵ para: (a) todos os 200 sinais de fala não anecoicos da base NBP, (b) todos os 32 sinais da base MARDY.

a base MARDY pode ser explicado pela distribuição espectral mais pobre, como mostrada na Fig. 3.4 (a), particularmente no intervalo entre 7 kHz e 8 kHz.

3.4.2 QAreverb com a base de dados NBP

Utilizando $\epsilon = 10^{-5}$, o valor do parâmetro γ de (3.1) foi obtido através da observação do coeficiente de correlação ρ entre o valor de Q e a nota subjetiva MOS correspondente a todos os sinais da NBP. Como pode ser observado na Fig. 3.6, o valor $\gamma = 0{,}3$ acarreta valor máximo para o coeficiente de correlação $\rho = 91\%$, para a base de dados NBP, o que requer os coeficientes do mapeamento não-linear de (3.7) $x_1 = 0{,}0017$, $x_2 = 0{,}0598$, $x_3 = 0{,}7014$ e $x_4 = 4{,}5387$ obtidos através de otimização numérica utilizando 102 sinais da base NBP. Nessa figura, para cada valor de γ avaliado uma, otimização de (x_1, x_2, x_3, x_4) foi feita, incluindo os casos especiais $\gamma = 0$ e $\gamma = 0{,}3$. A fim de minimizar o MSE entre Q_{MOS} ($\gamma = 0{,}3$) e as notas subjetivas dos sinais da base NBP, os coeficientes $\alpha = 1{,}00$ e $\beta = 1{,}85 \times 10^{-10}$ foram utilizados em (3.8), originando os resultados mostrados na Fig. 3.7.

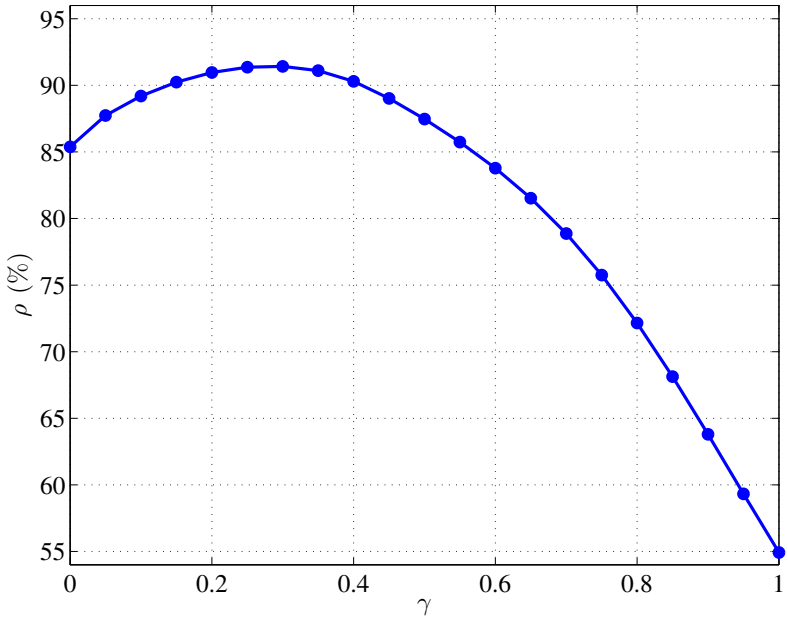


Figura 3.6: Influência do expoente γ da razão de energia no coeficiente de correlação ρ entre a nota objetiva Q_{MOS} e a nota subjetiva para a base NBP.

Para as partições "Natural" e "Artificial" da base NBP, a métrica Q_{MOS} foi determinada a partir das estimativas dos parâmetros T_{60}, σ_r^2 e E_{dr} feitas em duas versões da RIR referente a cada sinal: (i) RIR convolvida com o sinal anecoico para formar o sinal em questão; (ii) RIR estimada a partir da desconvolução de (3.3). Tanto para $\gamma = 0$ quanto para $\gamma = 0{,}3$, o coeficiente de correlação entre os dois resultados de Q_{MOS} foi de 99,9%, indicando que a RIR estimada por (3.3)), com

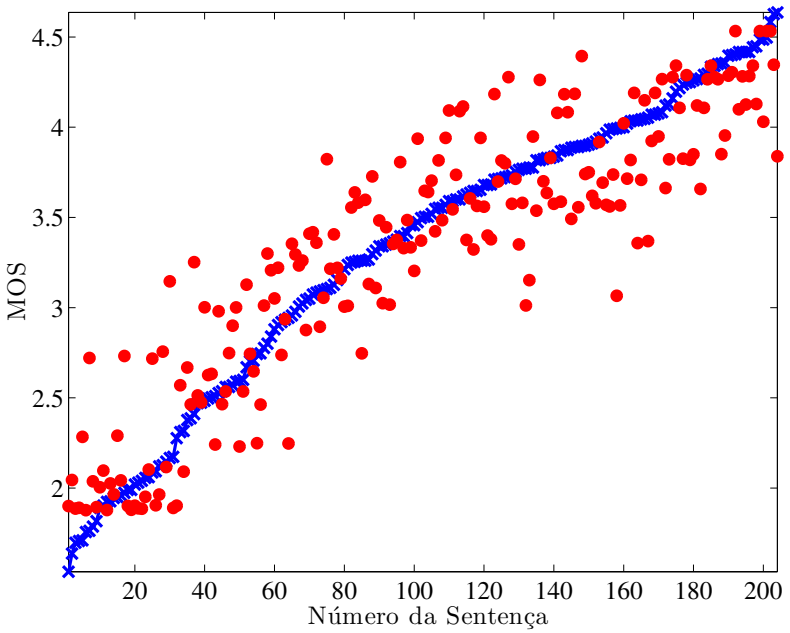


Figura 3.7: Notas objetivas Q_{MOS} (linha conectada com '×') e notas subjetivas MOS (pontos isolados marcados com '•') para os 204 sinais da base NBP, com $\gamma = 0{,}3$.

uma escolha adequada de ϵ, não causa impacto significativo no desempenho do sistema QAreverb.

A Tabela 3.3 mostra o coeficiente de correlação entre as notas subjetivas e notas objetivas determinadas por diversas técnicas encontradas na literatura para todas as partições da base NBP. As técnicas são: W-PESQ [37], ITU-T P.563 [39], R_{DT} [32, 40], SRMR [35] e Q_{MOS} com $\gamma = 0$ e $\gamma = 0{,}3$ Nesse conjunto, os algoritmos P.563 e SRMR são sem referência (dependem apenas do sinal degradado), enquanto os outros algoritmos são com referência, precisando também do sinal não degradado. Geralmente, algoritmos com referência tendem a ter um melhor desempenho em relação aos algoritmos sem referência. O impacto da largura de banda na avaliação de qualidade de sinais de fala reverberantes ainda é um problema em aberto na literatura associada. De qualquer modo, ambas as bases (NBP e MARDY) foram devidamente sub-amostradas para cada largura de banda do algoritmo em questão, 8 kHz no caso do algoritmo SRMR.

A Tabela 3.3 inclui o coeficiente de correlação para todos os algoritmos em questão, com e sem o mapeamento de terceira ordem de (3.7) otimizado para cada algoritmo utilizando a base NBP inteira, o que explica qualquer possível queda no valor do coeficiente de correlação nas três primeiras colunas ao se usar o mapeamento. Apesar de ambas as recomendações da ITU-T não serem originalmente para a avaliação de sinais reverberantes, o algoritmo W-PESQ apresentou um valor significativo para o coeficiente de correlação, especialmente ao se incorporar o mapeamento de terceira ordem. Para o caso sem referência, o algoritmo SRMR

representa o atual estado da arte, alcançando um coeficiente de correlação de 81%. A partir da Tabela 3.3 conclui-se que o sistema QAreverb, utilizando o mapeamento de terceira ordem e $\gamma = 0{,}3$, teve melhor desempenho do que os demais algoritmos, incluindo o caso particular de $\gamma = 0$, que corresponde à métrica original de Allen, alcançando um coeficiente de correlação de 91% com a base NBP completa.

Usando o teste-z de Fisher [50] para comparar as notas Q_{MOS} com $\gamma = 0$ e com $\gamma = 0{,}3$ na base NBP, obtém-se $p = 0{,}015$, o que estabelece um nível de confiança de 98,5% de que essas notas não são provenientes de uma mesma distribuição.

Tabela 3.3: Coeficiente de correlação ρ [sem/com mapeamento ótimo de terceira ordem descrito em (3.7)] entre as notas subjetivas e notas objetivas fornecidas por diferentes algoritmos de avaliação de qualidade para sinais de fala reverberantes aplicados à base NBP

Algoritmo Objetivo	Coeficiente de correlação (ρ) [%]			
	Artificial	Natural	Real	NBP inteira
W-PESQ	84/72	84/94	86/93	77/89
P.563	08/14	64/65	45/60	52/59
R_{DT}	68/69	75/80	43/43	59/61
SRMR	73/77	80/84	70/80	74/81
$Q_{\mathrm{MOS}}(\gamma = 0)$	90/89	92/92	86/88	85/85
$Q_{\mathrm{MOS}}(\gamma = 0{,}3)$	90/91	85/96	80/88	81/91

3.4.3 QAreverb com a base de dados MARDY

A base de dados MARDY foi utilizada neste trabalho para validar o desempenho do sistema QAreverb, visto que essa base é formada por sinais de fala reverberantes não utilizados no treinamento do sistema. Observando a Fig. 3.6, nota-se que os valores $\gamma = 0{,}3$ e o mapeamento ótimo ($x_1 = 0{,}0017$, $x_2 = 0{,}0598$, $x_3 = 0{,}7014$ e $x_4 = 4{,}5387$) para a base NBP também resultam em um alto valor para o coeficiente de correlação entre as notas objetivas Q_{MOS} e as notas subjetivas MOS para os sinais da base MARDY ($\rho = 95\%$). Esse alto valor de coeficiente de correlação em comparação com o obtido para a base NBP $\rho = 91\%$, pode ser explicado pelo fato de a base MARDY ter apenas 4 tipos de reverberação (painel reflexivo, painel absortivo e seus correspondentes desreverberados) e uma menor quantidade de sinais, o que leva a um processo mais facilmente modelável.

Como detalhado em [32, 40], o teste subjetivo feito na MARDY é composto de três diferentes experimentos referentes ao efeito da coloração (*coloration*), que consiste na distorção do espectro de sinais de fala causada pela resposta não-plana das primeiras reflexões [3] e é caracterizada nesta tese por σ_r^2; ao efeito da cauda reverberante (*reverberation tail effect*), que consiste no espalhamento do espectro e reduz a

inteligibilidade de sinais de fala [3] e é caracterizado nesta tese pelo T_{60}; e à qualidade geral de fala, todos utilizando a escala MOS. A Tabela 3.4 apresenta o coeficiente de correlação entre o valor MOS para esses três experimentos e as notas obtidas com os algoritmos W-PESQ, P.563, R_{DT}, SRMR e Q_{MOS} (com $\gamma = 0$ e $\gamma = 0{,}3$) com/sem o mapeamento ótimo de terceira ordem. A Tabela 3.4 também apresenta uma separação entre os sinais reverberantes e os sinais desreverberados através do algoritmo atraso-e-soma da base MARDY. A partir do resultado apresentado, mais uma vez pode ser observado que o sistema QAreverb com $\gamma = 0{,}3$ obteve um melhor desempenho do que os demais métodos, atingindo o maior valor do coeficiente de correlação para a coluna de qualidade geral de fala, sendo a avaliação de qualidade geral o foco principal da métrica de avaliação proposta nesta tese.

Tabela 3.4: Coeficiente de correlação ρ (sem/com mapeamento ótimo de terceira ordem descrito em (3.7)) entre as notas subjetivas e notas objetivas fornecidas por diferentes algoritmos de avaliação de qualidade para sinais de fala reverberantes aplicados à base MARDY

Algoritmo Objetivo	Coeficiente de correlação (ρ) [%]				
	Coloração	Efeito da cauda	Reverberante	Desreverberado	Qualidade Geral
W-PESQ	70/72	80/87	69/75	78/81	72/77
P.563	44/46	49/51	61/59	42/42	55/54
R_{DT}	59/61	59/60	70/69	51/52	64/64
SRMR	84/76	82/82	78/78	80/74	79/77
$Q_{\mathrm{MOS}}(\gamma = 0)$	90/90	97/97	91/91	93/93	91/92
$Q_{\mathrm{MOS}}(\gamma = 0{,}3)$	88/90	94/95	97/96	94/95	95/95

Como os testes subjetivos aplicados a cada base compreendem diferentes intervalos de degradação, os valores para os coeficientes do mapeamento linear de (3.8) são diferentes para cada base, como discutido na Seção 3.3, sendo $\alpha = 1{,}3314$ e $\beta = -1{,}4224$ para a base MARDY, levando ao resultado mostrado na Fig. 3.8.

A diferença estatística entre as notas objetivas Q_{MOS} obtidas com $\gamma = 0$ e $\gamma = 0{,}3$ para a base MARDY foi verificada a partir do teste-z de Fisher, com um resultado $p = 0{,}30$, o que significa um nível de confiança de 70% para a hipótese de distribuições distintas. O menor valor de confiança alcançado em relação ao obtido para a base NBP, pode ser atribuído ao menor número de sinais, 32 contra 204, contidos na base MARDY.

3.5 Conclusões

Neste capítulo, foi apresentada a base de dados NBP, composta de 204 sinais de fala reverberantes, cada um avaliado subjetivamente por 30 ouvintes. Um sistema

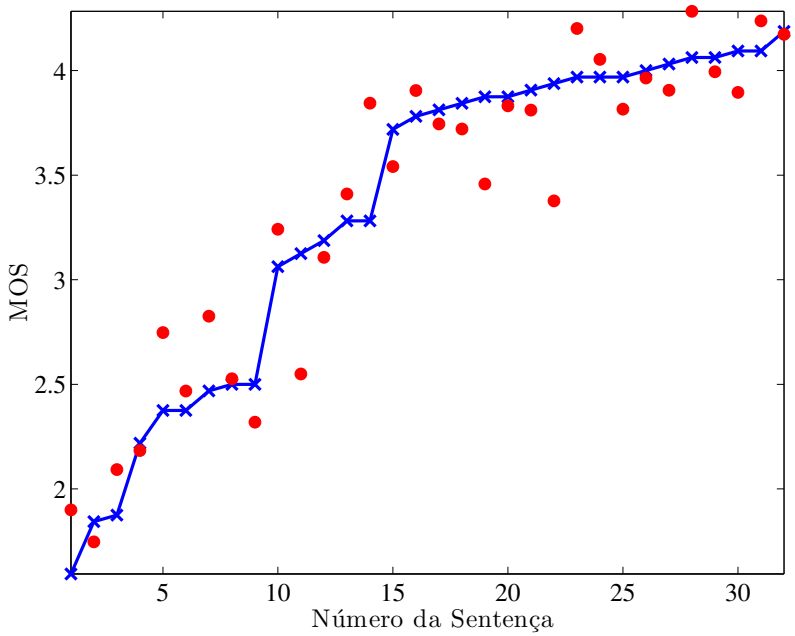


Figura 3.8: Notas objetivas Q_{MOS} (linha conectada com '$\times$') e notas subjetivas MOS (pontos isolados marcados com '$\bullet$') para os 32 sinais da base MARDY, com $\gamma = 0{,}3$.

completo para a avaliação de qualidade perceptual de sinais de fala afetados por reverberação foi descrito. Esse sistema teve como base a métrica de Allen, na qual foi incorporada a razão de energia direta sobre reverberante E_{dr}. Um procedimento completo para o treinamento de um sistema de avaliação de qualidade de sinais reverberantes foi detalhado, utilizando a base NBP. O sistema foi testado a partir de duas bases independentes NBP e MARDY, atingindo, respectivamente, valores de 91% e 95% para os coeficientes de correlação entre as notas objetivas obtidas com o sistema e as notas subjetivas correspondentes, melhorando a proposta original de Allen e superando o desempenho de outros métodos, encontrados na literatura, para avaliação de qualidade de sinais de fala.

Capítulo 4

Estimador de T_{60} sem referência

4.1 Introdução

Este capítulo aborda o problema de se estimar de maneira cega o tempo de reverberação de um ambiente a partir de um sinal de fala reverberante. Esse tipo de estimação também é conhecido como sem referência, uma vez que o sinal original não é utilizado. A estimação cega é muito importante para o monitoramento de características de um dado sistema de comunicação no qual o sinal sem degradação (enviado) não está disponível no receptor.

Trabalhos iniciais na estimação cega de T_{60} incluem as referências [16, 51], nas quais os autores modelam o processo de decaimento como uma função exponencial, cuja constante de tempo é estimada a partir do sinal completo. Mais tarde, Vieira [52] restringiu a modelagem do processo de decaimento às regiões intituladas regiões de decaimento livre (*free-decay regions* - FDRs), que são porções do sinal nas quais a energia sonora decresce em diversos segmentos consecutivos. Ao limitar-se a essas regiões, consegue uma melhor modelagem, resultando em estimativas mais acuradas de T_{60}. Em [53], Vieira incorporou ao seu algoritmo um modelo modificado de decaimento da energia proposto por [21], que considera presença de ruído aditivo, tornando seu algoritmo mais robusto a ruídos de medição. Outros trabalhos relacionados:

- Wu e Wang [3] propõem um método que utiliza um modelo de tempo de reverberação baseado no *pitch* e restringe a análise a um pequeno intervalo de possíveis valores para T_{60}.

- Wen et al. [54] propõem um método que requer um mapeamento quadrático altamente dependente do algoritmo.

- Löllmann e Vary [55] incorporam um estágio para redução de ruído ao método descrito em [16], utilizando todo o sinal, o que resulta em um processo de

estimação com alta variância.

Apesar de a restrição da estimação de T_{60} às FDRs resultar em um procedimento mais acurado, ela também obriga a utilização de sinais longos (por exemplo 40 s, como utilizado em [52, 53]) que contenham diversas alternâncias entre atividade sonora e silêncio, a fim de obter uma quantidade satisfatória de estimativas. O algoritmo proposto, que também utiliza a estimação somente nas FDRs, reduz a necessidade de sinais longos ao fazer uma decomposição espectral do sinal reverberante, com uma abordagem semelhante à usada em [32, 56]. Feito isso, as FDRs são localizadas em cada sub-banda e o tempo de reverberação é estimado para cada FDR localizada, o que resulta em uma grande quantidade de estimativas parciais de T_{60}, inclusive para sinais relativamente curtos, tornando o algoritmo final adequado para utilização em tempo real.

O algoritmo proposto para a estimação cega de T_{60} é descrito na Seção 4.2. A Seção 4.3 discute alguns aspectos práticos do treinamento do sistema, além de apresentar a validação do desempenho do algoritmo proposto para a estimação cega de T_{60}, utilizando as bases MARDY e NBP. Por fim, a Seção 4.4 apresenta um resumo do que foi discutido neste capítulo.

4.2 Estimação cega de T_{60}

O algoritmo proposto para a estimação de T_{60} é composto por quatro etapas:

- Representação tempo-frequência do sinal reverberante $s_r(n)$.
- Localização das FDRs em cada sub-banda.
- Estimação do tempo de reverberação para todas as FDRs detectadas.
- Análise estatística das estimativas do tempo de reverberação para gerar a estimativa final do T_{60}.

4.2.1 Representação tempo-frequência

Neste estágio inicial, o sinal reverberante de fala $s_r(n)$ é dividido em segmentos utilizando uma função de janelamento $w(n)$ de tamanho M amostras. A DFT é aplicada em cada segmento, gerando a representação tempo-frequência

$$S_r(k,l) = \mathrm{DFT}[w(n)s_r(n)], \tag{4.1}$$

para $k = 0, 1, \ldots, (K-1)$, $l = 0, 1, \ldots, (L-1)$ e $n = l(M-V), l(M-V)+1, \ldots, l(M-V)+M-1$, em que K é o comprimento da DFT, L é o número total

de segmentos e V é o número de amostras de sobreposição entre dois segmentos consecutivos.

Como a maior parte da energia de sinais de fala se concentra na faixa de frequências $0 \leq f \leq 4$ kHz, todas as análises subsequentes são restringidas a valores de k tais que $0 \leq \frac{F_s k}{K} \leq 4$ kHz ($F_s \geq 8$ kHz), de forma a garantir estimativas de T_{60} menos suscetíveis à presença de ruído de fundo.

4.2.2 Detecção de FDR em uma dada sub-banda

Como mencionado na Seção 4.1, as FDRs são caracterizadas por uma queda de energia em segmentos consecutivos e, com o objetivo de aumentar a quantidade de FDRs localizadas em um dado sinal, este último é dividido em K bandas com frequência central $\frac{F_s k}{K}$ Hz.

As FDRs devem ser localizadas em cada sub-banda de maneira independente, uma vez que cada componente espectral apresenta um padrão distinto de energia [57]. Seja a energia da k-ésima sub-banda do l-ésimo segmento do sinal $s_r(n)$, dada por

$$E(k,l) = |S_r(k,l)|^2. \tag{4.2}$$

A busca pelas FDRs é feita no índice $l = 0, 1, \ldots, (L-1)$ para cada sub-banda $k = 0, 1, \ldots, (K-1)$.

Ao se estender o critério de Vieira [52] para sub-bandas, uma FDR na sub-banda k pode ser caracterizada por um decréscimo no valor de $E(k,l)$ em pelo menos 500 ms ao longo de l. Utilizando-se M amostras por segmento com sobreposição de V amostras por segmento, esse intervalo de 500 ms é traduzido em

$$L_{\text{lim}} = \frac{0{,}500 F_s}{M - V} \tag{4.3}$$

segmentos consecutivos com energia decrescente. Neste trabalho $M = 0{,}05 F_s$ e $V = 0{,}25M$, o que resulta em $L_{\text{lim}} \approx 13$ segmentos consecutivos. A escolha dos parâmetros M e V será detalhada na Seção 4.3.1.

No sistema proposto, caso não seja encontrada pelo menos uma FDR em uma dada sub-banda, L_{lim} é reduzido em 1 segmento iterativamente até que se encontre pelo menos uma FDR ou $L_{\text{lim}} < 3$ segmentos. Esse limite $L_{\text{lim}} = 3$ foi determinado de maneira experimental e garante que pelo menos uma FDR por sub-banda será encontrada. Esse decréscimo iterativo garante uma quantidade mínima de dados significativos que será utilizada nas próximas etapas do sistema.

O processo de detecção de FDRs em um sinal composto por duas sentenças é exemplificado na Fig. 4.1, na qual as linhas horizontais de cor escura indicam as FDRs em cada sub-banda. A partir desta figura é possível observar os padrões de

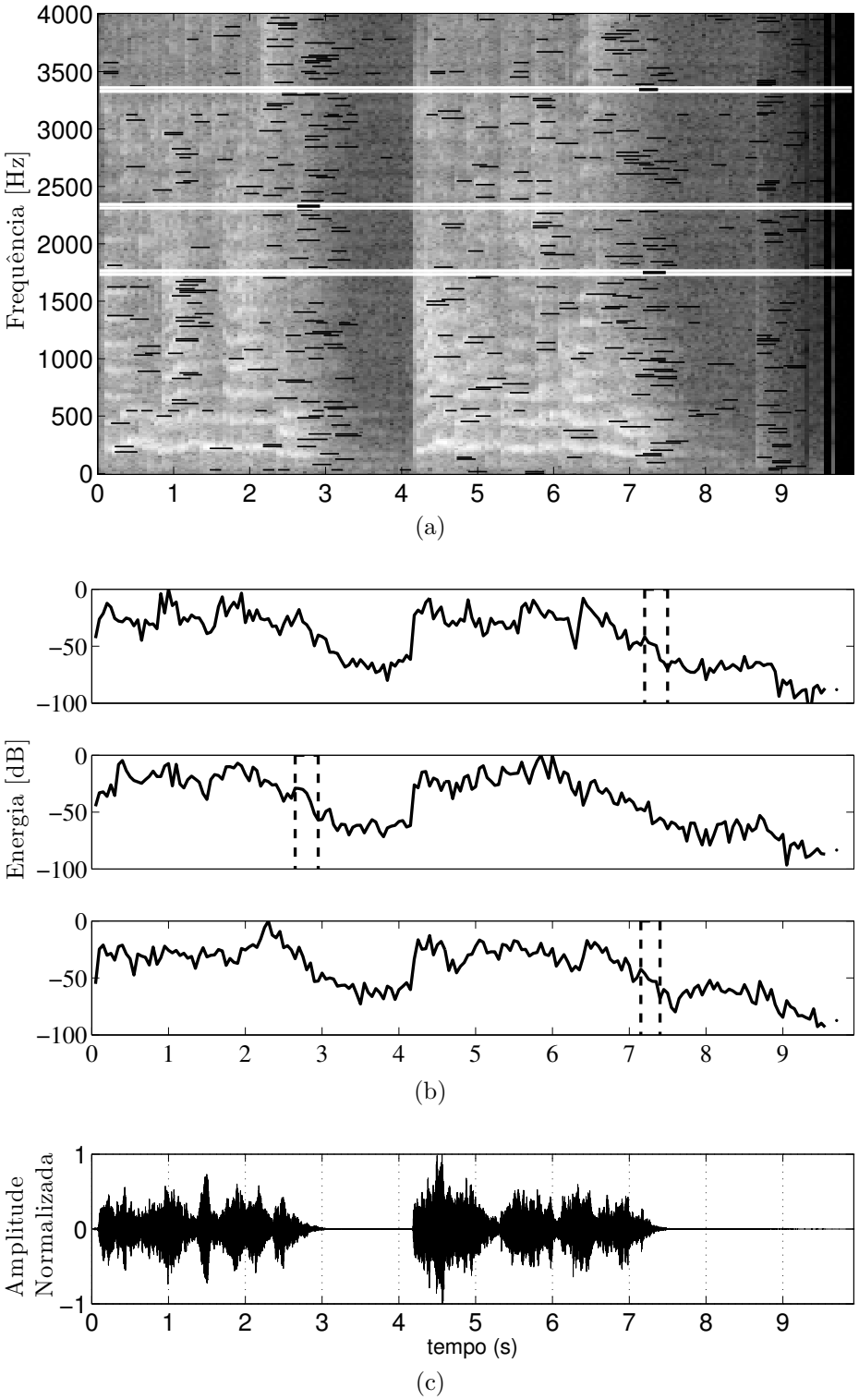


Figura 4.1: Caracterização de FDRs em sub-bandas: (a) Espectrograma mostrando todas as sub-bandas e respectivas FDRs (utilizando $M = 0,05\,F_s$ e $V = M/4$) como linhas finas de cor escura; (b) Três sub-bandas [identificadas por linhas horizontais de cor branca em (a)], com frequências centrais em 1750, 2330, e 3340 Hz, respectivamente, mostrando as FDRs correspondentes com linhas verticais tracejadas; (c) Sinal de fala com duas sentenças.

FDR distintos para cada sub-banda, com uma notável concentração das FDRs no início dos intervalos de silêncio, nos quais o processo de reverberação em banda completa é dominante.

4.2.3 Estimação de tempo de reverberação em sub-bandas

Algoritmos encontrados na literatura [18, 21, 23] estimam o T_{60} a partir de uma regressão linear para a EDC. O fator primordial para esses algoritmos é encontrar o intervalo de tempo $n_1 \leq n \leq n_2$ que resulte em uma regressão linear confiável. É comum escolher n_1 de modo que $\xi(n_1) = -5$ dB, enquanto n_2 é escolhido de modo a minimizar o MSE entre a EDC e a regressão linear resultante.

Em geral, os algoritmos descritos em [21, 23] tendem a ter um comportamento robusto na presença de ruído aditivo não correlacionado, porém exigem uma quantidade significativamente maior de pontos na EDC para gerar uma regressão linear confiável, em comparação ao método de Schroeder [18], o que inviabiliza a utilização de ambos no sistema proposto, já que existem poucos pontos devido ao processo de determinação de $S(k, l)$.

Assim, uma extensão do método de Schroeder para sub-bandas é empregada. Seja a SEDC (EDC em sub-bandas) dada por

$$\xi^s(k,l) = 10\log_{10}\left(\frac{\sum_{\lambda=l}^{\bar{L}-1} E(k,\lambda)}{\sum_{\lambda=0}^{\bar{L}-1} E(k,\lambda)}\right), \tag{4.4}$$

para $l = 0, 1, \ldots, (\bar{L}-1)$, em que $\bar{L}$ é o número de segmentos que compõem uma dada FDR em sub-banda. O tempo de reverberação em sub-banda T_{60}^s é estimado em cada FDR de modo semelhante ao descrito anteriormente, isto é, faz-se a regressão linear no intervalo $l_1 \leq l \leq l_2$, em que $\xi^s(k, l_2) - \xi^s(k, l_1) = 60$ dB.

Utilizando sinais reais de fala, é possível que uma queda de 60 dB nas SEDCs não seja observada. Nesses casos, a regressão linear de Schroeder considera um intervalo de atenuação inferior a 60 dB e, com isso, o T_{60}^s deve ser extrapolado a partir do tempo de reverberação calculado. Ao lidar com segmentos ao invés de amostras, a resolução dentro do intervalo $[l_1, l_2]$ diminui devido ao efeito de janelamento, aumentando a variância da estimativa de T_{60}^s de maneira significativa, particularmente quando l_2 é próximo de l_1. Para minimizar esse efeito, se a escolha de l_2 que minimize o MSE entre a SEDC e a regressão linear for tal que $\xi^s(k, l_2) - \xi^s(k, l_1) < 10$ dB, uma nova regressão linear é calculada, sempre que possível, de modo que $\xi^s(k, l_2) = -65$, -45, -25 ou -15 dB, nessa ordem de preferência. Como $\xi^s(k, l_1) = -5$ dB, os novos valores utilizados levam aos tempos de reverberação T_{60}, T_{40}, T_{20} e T_{10}, res-

pectivamente, como definidos em [19]. Assumindo um decaimento de energia linear, o tempo de reverberação na escala desejada pode ser obtido a partir de qualquer outro tempo, por exemplo, pode-se obter uma estimativa para o T_{60} a partir da estimativa do T_{20}.

4.2.4 Análise estatística do tempo de reverberação em sub-bandas

Partindo do pressuposto que foram encontradas R_k FDRs na k-ésima sub-banda, cada estimativa parcial do tempo de reverberação pode ser denotado por $T_{60}^s(r,k)$, para $r = 0, 1, \ldots, (R_k - 1)$. O objetivo final do algoritmo proposto é estimar o tempo de reverberação em banda completa T_{60} a partir das estimativas parciais $T_{60}^s(r,k)$.

Ratnam et al. [16] propõe diversas estratégias para se remover estimativas parciais espúrias, o que não precisa ser feito no algoritmo proposto neste trabalho, uma vez que estima-se o tempo de reverberação apenas nas FDRs. Vieira [52, 53], por outro lado, obtém a estimativa final do T_{60} a partir do pico do histograma dos tempos de reverberação parciais, processo esse altamente dependente da resolução escolhida para o histograma.

Propõe-se, então, que primeiramente se calcule a mediana $\hat{T}_{60}^s(k)$ das R_k estimativas $T_{60}^s(r,k)$ para todas as sub-bandas. Depois, calcula-se uma primeira estimativa $\bar{T}_{60}^s$ para o tempo de reverberação em banda completa como a mediana de $\hat{T}_{60}^s(k)$, o que evita valores extremos. Na verdade, o operador mediana elimina estimativas em sub-banda com valores relativamente pequenos (que não afetam significativamente a dinâmica em banda completa) e grandes (que podem carregar grandes erros de estimação), levando a uma estimativa parcial que aparentemente represente a dinâmica do tempo de reverberação em banda completa de uma maneira consistente.

A relação entre o tempo de reverberação em sub-bandas e o tempo de reverberação em banda completa é algo bastante difícil de modelar e constitui um problema ainda aberto na literatura [54, 56]. As estimativas obtidas com o algoritmo proposto, apesar de altamente correlacionadas com os valores de referência da base NBP, variam em um intervalo dinâmico diferente, o que leva à necessidade de uma função de mapeamento, dada por

$$\tilde{T}_{60}^s = \alpha_{nr} \bar{T}_{60}^s + \beta_{nr}, \tag{4.5}$$

com α_{nr} e β_{nr} obtidos no estágio de treinamento do algoritmo. Esse mapeamento diminui o MSE entre as estimativas obtidas e os valores de referência, sem afetar a correlação entre eles.

4.3 Aspectos práticos

4.3.1 Ajuste do algoritmo de estimação do tempo de reverberação

A base NBP foi dividida em duas sub-bases A_a e A_v contendo 100 sinais reverberantes em cada. Os sinais anecoicos não foram incluídos em nenhuma das sub-bases, uma vez que seus valores de T_{60} são da ordem do decaimento existente no término de palavras faladas, implicando resultados inconsistentes para a estimação desses tempos de reverberação devido a ausência de pontos suficientes. A sub-base A_a foi utilizada no ajuste dos parâmetros do algoritmo proposto para a estimação de T_{60}, enquanto a sub-base A_v e a base MARDY foram utilizadas para a validação do desempenho geral do algoritmo.

Os parâmetros considerados na análise são a duração do segmento ($W = \frac{M}{F_s}$), a sobreposição percentual entre segmentos adjacentes $v = \frac{V}{M} \times 100\%$ e o número K de *bins* da DFT com frequência central dentro do intervalo entre 0 e 4 kHz. O desempenho foi aferido através do coeficiente de correlação entre $\bar{T}_{60}^s$ e o valor de T_{60} estimado pelo algoritmo descrito em [23], denotado T_{60} teórico, como mostrado na Tabela 4.1, para $v = 25\%$ e $K = 1024$. Os valores $v = \{0\%, 50\%, 75\%\}$ e $K = \{512, 2048\}$ foram testados em outros experimentos, sem qualquer aumento do desempenho do sistema. Com base nos resultados apresentados nesta tabela, conclui-se que $W = 50$ ms gera o melhor resultado (em negrito), com um coeficiente de correlação de aproximadamente 92% para A_a.

Utilizando a sub-base de treinamento A_a, os parâmetros do mapeamento de (4.5) escolhidos são $\alpha_{nr} = 3{,}4$ e $\beta_{nr} = -1170$ ms, que minimizam o MSE entre $\tilde{T}_{60}^s$ e T_{60}, sem afetar o coeficiente de correlação.

4.3.2 Validação do algoritmo de estimação do tempo de reverberação

O desempenho do algoritmo para A_v também é apresentado na Tabela 4.1, em que o coeficiente de correlação para os parâmetros escolhidos é de 91%. A Fig. 4.2 mostra os valores de $\tilde{T}_{60}^s$ e T_{60} para toda a base NBP, utilizando a configuração escolhida, mostrando a capacidade de o algoritmo proposto em fornecer uma estimativa confiável para um amplo intervalo de tempos de reverberação.

A Fig. 4.3 mostra os valores de $\tilde{T}_{60}^s$ e T_{60} para toda a base MARDY, em que o coeficiente de correlação entre os dois estimadores é de aproximadamente 97%. O significativo aumento no coeficiente de correlação em relação ao obtido para a base NBP pode ser explicado pelo reduzido escopo coberto pela base MARDY em relação

Tabela 4.1: Coeficiente de correlação entre $\bar{T}_{60}^{s}$ e T_{60} teórico para a base NBP com distintos valores de W, em que $v = 25\%$ e $K = 1024$

W [**ms**]	A_a	A_v
30	86,4	84,4
35	88,7	88,1
40	89,3	88,2
45	92,0	90,1
50	**92,1**	**91,0**
55	91,1	89,4
60	89,6	88,8
65	91,5	90,7
70	89,9	89,3
75	89,8	86,7
80	88,0	85,0
85	86,6	85,4
90	85,2	85,3
95	84,6	81,2
100	84,2	84,0

ao escopo coberto (três diferentes tipos de reverberação, maior intervalo de T_{60} e σ_r^2 etc) pela base NBP.

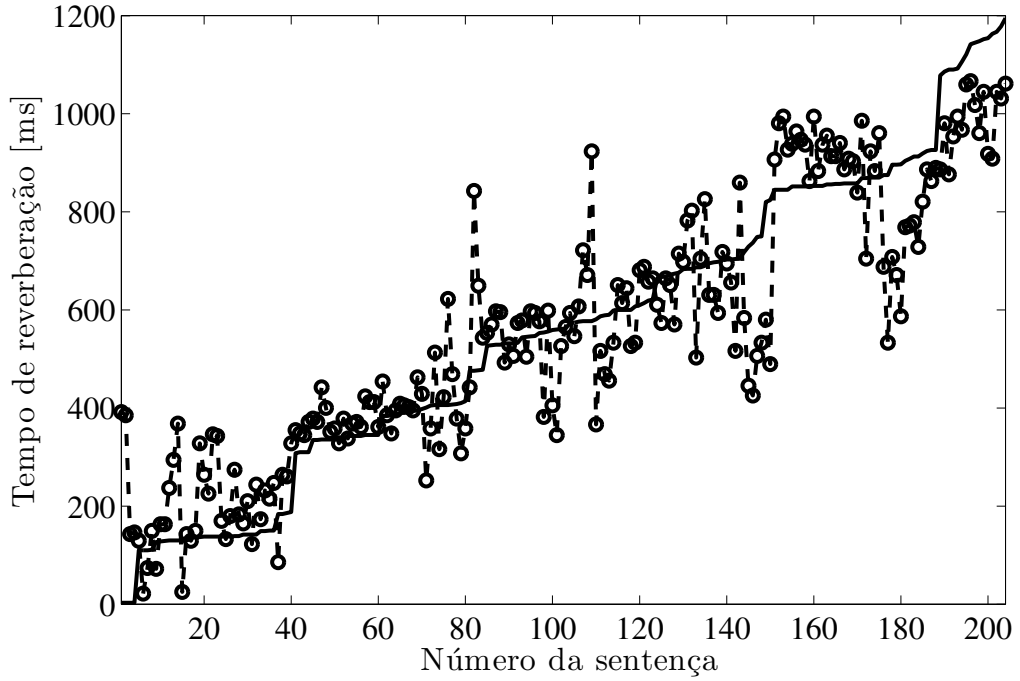


Figura 4.2: Tempos de reverberação estimados utilizando o algoritmo cego proposto (linha tracejada) e o método referência não cego (linha cheia) para todos os 204 sinais da base NBP.

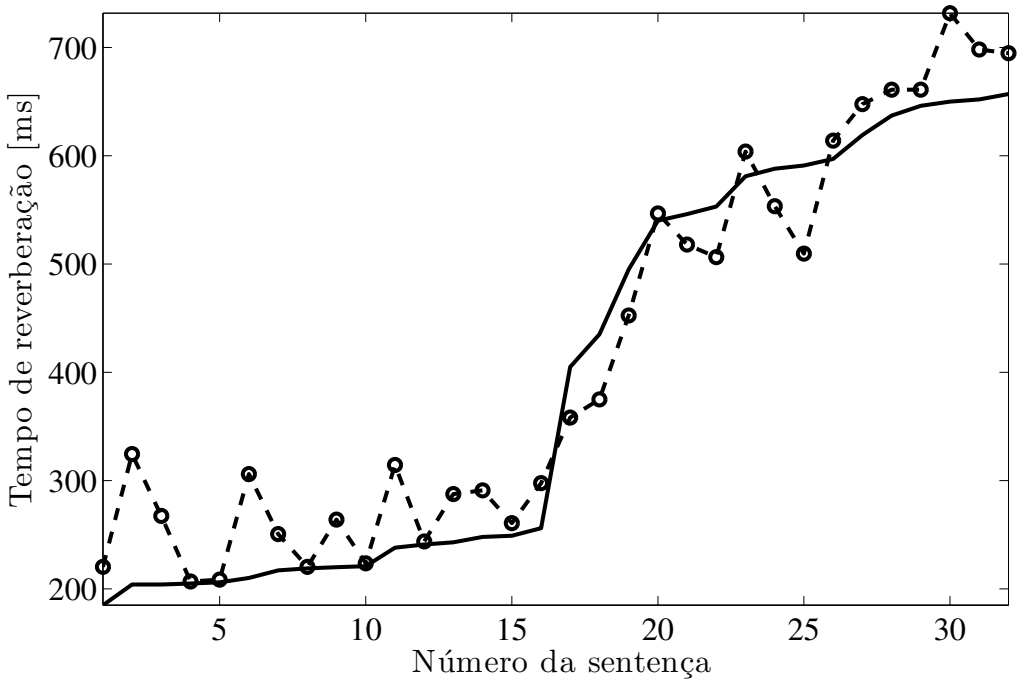


Figura 4.3: Tempos de reverberação estimados utilizando o algoritmo cego proposto (linha tracejada) e o método referência não cego (linha cheia) para todos os 32 sinais da base MARDY.

4.3.3 Comparação com outros métodos

A Tabela 4.2 mostra o coeficiente de correlação ρ e o desvio padrão σ entre os valores de T_{60} teórico e estimado para as bases NBP e MARDY utilizando os algoritmos descritos nas referências [16, 53]. Essa tabela também inclui resultados fornecidos por diversos algoritmos para estimação de qualidade percebida de sinais de fala, que, em alguns casos, são altamente correlacionados com o tempo de reverberação. A partir dessa tabela, conclui-se que o algoritmo proposto obtém a maior correlação e o menor desvio padrão dentre os algoritmos em questão, tanto para a base NBP quanto para a base MARDY.

Tabela 4.2: Coeficiente de correlação (ρ) e desvio padrão (σ) entre T_{60} teórico e estimado para diversos algoritmos de estimação de tempo de reverberação ou de estimação de qualidade para as bases NBP e MARDY

Algoritmo de estimação	**NBP**		**MARDY**	
	ρ [%]	σ [ms]	ρ [%]	σ [ms]
Ratnam et al. [16]	55	254	50	178
Vieira [53]	64	234	26	290
R_{DT} [32]	58	248	57	190
SRMR [56]	72	213	84	105
ITU-T W-PESQ [37]	76	197	78	118
ITU-T P.563 [39]	29	293	25	198
Algoritmo proposto	**91**	**124**	**97**	**46**

4.4 Conclusões

Este capítulo descreveu uma técnica para estimação cega do tempo de reverberação T_{60} a partir de sinais de fala reverberantes. A técnica proposta inclui quatro estágios simples que utilizam o sinal de fala particionado em segmentos de mesmo tamanho. O desempenho do método proposto foi aferido a partir de duas bases independentes contendo sinais de fala reverberantes, proporcionando alta correlação e baixo desvio padrão em comparação a outros métodos encontrados na literatura.

Os resultados indicam que a técnica proposta tem a capacidade de monitorar o efeito da reverberação em sistemas práticos de comunicação em que o sinal de referência (enviado) não está disponível para o avaliador (receptor).

Capítulo 5

QAreverb cego

5.1 Introdução

A técnica de avaliação de qualidade descrita neste capítulo tem por objetivo resolver o problema de estimar a qualidade percebida de sinais de fala utilizando apenas o sinal de fala reverberante obtido por um único microfone. A técnica descrita é sem referência e não intrusiva, uma vez que não necessita do sinal original (sem reverberação) e não requer qualquer interferência no sistema de comunicação a fim de se obter o sinal degradado, sendo também chamada de uma técnica cega.

A técnica descrita neste capítulo tem como base o algoritmo não cego descrito no Capítulo 3, que requer as métricas T_{60}, σ_r^2 e E_{dr}, definidas em (2.2), (2.4) e (2.5), respectivamente. Originalmente essas três métricas são obtidas utilizando-se a RIR. A fim de se obter estimativas cegas dessas três métricas, foram selecionadas, a partir da literatura, 3 abordagens para estimar T_{60}[16, 53, 54], além do abordagem descrita no Capítulo 4, uma abordagem para estimar σ_r^2 [58] e 2 abordagens para estimar E_{dr}, das quais uma foi obtida da literatura [56] e a outra é descrita neste capítulo. O objetivo principal deste trabalho é obter a combinação desses algoritmos que tenha o maior coeficiente de correlação entre a nota cega equivalente a Q (3.1) e a nota MOS.

O esquema cego resultante pode ser utilizado para avaliar técnicas de desreverberação, evitando o uso de métodos subjetivos custosos e difíceis de reproduzir. Esse capítulo está organizado da seguinte maneira: a Seção 5.2 descreve os algoritmos para estimar T_{60} propostos nas referências [16, 53, 54]; a Seção 5.3 apresenta um algoritmo que emula o método de estimação de σ_r^2 proposto em [56]; a Seção 5.4 descreve os dois algoritmos para estimação de E_{dr}; a Seção 5.5 detalha os resultados experimentais e compara as estimativas cegas com seus correspondentes não cegos, além de avaliar o desempenho do sistema cego como um todo utilizando as bases NBP e MARDY descritas nos Capítulos 2 e 3. A Seção 5.6 resume as principais

contribuições deste capítulo.

5.2 Estimadores cegos de T_{60}

5.2.1 Estimador proposto para T_{60}

O estimador descrito no Capítulo 4 é referenciado neste capítulo como o estimador proposto para T_{60}, sendo representado por $\hat{T}_{60}^{(P)}$.

5.2.2 Estimador de Ratnam para T_{60}

Ratnam et al. [16] propõem um algoritmo para estimar o tempo de reverberação T_{60} com base em um ruído branco gaussiano amortecido exponencialmente. Esse algoritmo estima o tempo de reverberação através da máxima verossimilhança.

O decaimento sonoro é modelado por

$$s_r(n) = a^n x(n), \tag{5.1}$$

em que $x(n)$ é uma sequência aleatória com distribuição normal $\mathcal{N}(0, \sigma)$, $a = e^{-1/\tau}$ e τ é a taxa de decaimento que é relacionada ao T_{60} através de

$$T_{60} = 6{,}901\tau. \tag{5.2}$$

Com o objetivo de estimar τ, considera-se um número finito de observações, $n = 0, \ldots, M-1$, em que M é o intervalo de estimação ou tamanho da janela de estimação em amostras. Por simplicidade de notação, o vetor M-dimensional de $s_r(n)$ é denotado por $\mathbf{s_r}$. Dessa forma, a função de verossimilhança de $\mathbf{s_r}$, parametrizada por a e σ é dada por

$$L(\mathbf{s_r}; a, \sigma) = \left(\frac{1}{2\pi a^{M-1}\sigma^2} \right)^{M/2} \times \exp\left(- \frac{\sum_{n=0}^{(M-1)} \left(a^{-n} s_r(n)\right)^2}{2\sigma^2} \right), \tag{5.3}$$

em que a e σ são os dois parâmetros desconhecidos a serem estimados utilizando a observação $\mathbf{s_r}$.

Essa estimação é feita através da abordagem da máxima verossimilhança, de forma que $\dfrac{\partial L(\mathbf{s_r}; a, \sigma)}{\partial a} = \dfrac{\partial \ln L(\mathbf{s_r}; a, \sigma)}{\partial a} = 0$ e $\dfrac{\partial L(\mathbf{s_r}; a, \sigma)}{\partial \sigma} = \dfrac{\partial \ln L(\mathbf{s_r}; a, \sigma)}{\partial \sigma} = 0$.

Assim,

$$-\frac{M(M-1)}{2a}+\frac{1}{a\sigma^2}\sum_{n=0}^{M-1}(na^{-2n}s_r(n)^2) \;=\; 0 \tag{5.4}$$

$$\sigma^2=\frac{1}{M}\sum_{n=0}^{M-1}\left(a^{-n}s_r(n)\right)^2. \tag{5.5}$$

(5.4) não possui solução direta e uma combinação do método de bisseção com o método de Newton-Raphson é empregada para se obter uma estimativa de a. Em [16], o sinal de fala reverberante $s_r(n)$ é dividido em L segmentos de tamanho M amostras com $V = M - 1$ amostras de sobreposição, levando a uma grande quantidade de estimativas, dentre as quais muitas estarão erradas, já que o modelo da (5.1) só é válido para as FDRs. Uma das estratégias sugeridas em [16], para eliminar essas estimativas erradas, é fazer com que a cauda esquerda da densidade de probabilidade da estimativa de a ocupe um percentil ϕ. ϕ é a razão entre soma da duração de todas as FDRs e a duração total de $s_r(n)$.

Neste trabalho, a implementação do método proposto por Ratnam et al. [16] é feita com $M = 0{,}190F_s$ amostras, $V = 0{,}165F_s$ amostras e, ao invés de utilizar a combinação entre o método da bisseção e o método de Newton-Raphson, utilizou-se o método de *golden section search* [59]. A estimativa para T_{60} através desse método é denotada por $\hat{T}_{60}^{(R)}$.

5.2.3 Estimador de Vieira para T_{60}

O algoritmo cego proposto por Vieira [53] estima T_{60} nas FDRs. Esse algoritmo consiste em: encontrar as FDRs; obter uma estimativa de T_{60} para cada FDR; obter uma única estimativa para o sinal $s_r(n)$ através do histograma das estimativas parciais.

O sinal $s_r(n)$ é dividido em L segmentos de tamanho M amostras, com sobreposição $V = 0$. A energia do l-ésimo segmento é

$$E_l=\sum_{n=lM}^{(l+1)(M-1)} s_r^2(n). \tag{5.6}$$

e a l-ésima diferença de energia entre segmentos é dada por

$$d_l = E_l - E_{l-1}. \tag{5.7}$$

Caso t_{hr} segmentos consecutivos apresentem $d_l \leq 0$, essa região é dita FDR. Neste trabalho, é usado $M = 0{,}1F_s$ amostras e $t_{hr} = 5$ segmentos consecutivos.

Para estimar T_{60} de cada região, o algoritmo original de Vieira [53] aplica o método de Xiang [21] em cada FDR. Aqui, o algoritmo de Xiang foi substituído pelo algoritmo de Karjalainen et al. [23, 24], que apresentou resultados preliminares melhores do que o algoritmo de Xiang.

A estimativa final de Viera $\hat{T}_{60}^{(V)}$ é a moda extraída do histograma das estimativas de T_{60} para todas as FDRs detectadas.

É importante ressaltar que o método de Vieira [53] foi validado em um sinal de fala de 60 segundos de duração, a fim de gerar estimativas suficientes para o histograma. Nas bases abordadas por este trabalho, a duração média dos sinais é uma ordem de grandeza menor do que a duração do sinal utilizado por Vieira em [53].

5.2.4 Estimador de Wen para T_{60}

O algoritmo cego para estimar T_{60} baseado na distribuição da taxa de decaimento de sinais de fala reverberantes proposto por Wen et al. [54] utiliza um modelo similar ao apresentado em (5.1). O modelo em tempo contínuo para a RIR utilizado é tal que

$$h(t) = b(t)e^{-\tau t}, \tag{5.8}$$

em que $b(t)$ é um processo estacionário de ruído branco Gaussiano de média zero e τ é uma constante de decaimento, relacionada com o tempo de reverberação através de (5.2).

O modelo de decaimento do ambiente pode ser expresso pela variância σ_b^2 de $b(t)$ e pela taxa de decaimento $\lambda = -2\tau$, de modo que

$$\mathbf{E}[h^2(t)] = \sigma_b^2 \mathrm{e}^{\lambda t}, \tag{5.9}$$

em que $\mathbf{E}[\cdot]$ denota o operador valor esperado.

Assim, podemos reescrever (5.9) como

$$\lambda = \frac{1}{t} \ln \left(\frac{\mathbf{E}[h^2(t)]}{\sigma_b^2} \right). \tag{5.10}$$

Esse modelo pode ser estendido para um modelo com decaimentos dependentes da frequência, de modo que

$$\lambda(f) = \frac{1}{t} \left(\ln \tilde{H}(f,t) - \ln P(f) \right), \tag{5.11}$$

em que $\ln \tilde{H}(k,t)$ é a envoltória de energia da RIR no tempo t e na frequência f, $\lambda(f)$ é a taxa de decaimento na frequência f e $P(f)$ é a densidade espectral de potência

inicial. $\lambda(f)$ pode ser estimada através de uma regressão linear de $\ln \tilde{H}(k,t)$.

Apesar de o modelo se referir a RIR, o sinal observado é o sinal de fala reverberante, modelado como resultado da convolução linear entre a RIR e o sinal de fala anecoico. Segundo [54], a taxa de decaimento do ambiente é corretamente estimada nas FDRs.

A representação tempo-frequência é gerada utilizando-se uma versão discreta de (5.11) através da transformada de Fourier em tempo curto (STFT - *short-time Fourier transform*), utilizando uma janela de Hamming de tamanho M amostras com sobreposição de V amostras. Neste trabalho $M = 0{,}016F_s$ amostras, $V = 0{,}012F_s$ (equivalente a 75% de sobreposição) e $K = 512$ *bins* (tamanho, em amostras, da DFT).

Para o k-ésimo *bin* obtém-se uma estimativa para $\lambda(k)$ a cada 20 amostras (equivalentes a 20 segmentos no tempo) através de uma regressão linear do tipo *least squares*. Todas as estimativas de $\lambda(k)$ para todos os K *bins* são utilizadas para gerar um único histograma. A parte positiva desse histograma é descartada e a parte negativa é espelhada para o lado positivo. Calcula-se, então, a variância σ_λ^2 desse histograma modificado. Essa variância está relacionada com a estimativa final de λ da seguinte forma:

$$\hat{\lambda} = \gamma_2(\sigma_\lambda^2)^2 + \gamma_1\sigma_\lambda^2 + \gamma_0, \tag{5.12}$$

em que γ_0, γ_1 e γ_2 são os coeficientes de um mapeamento de segunda ordem obtidos na etapa de treinamento do algoritmo. Sendo assim, a estimativa final de Wen $\hat{T}_{60}^{(W)}$ é dada por

$$\hat{T}_{60}^{(W)} = \frac{3\ln 10}{\hat{\lambda}/2}. \tag{5.13}$$

5.3 Estimador de Habets para σ_r^2

Habets et al.[58] propõem um algoritmo para estimar a variância espectral da reverberação tardia. Esse algoritmo é baseado na representação tempo-frequência do sinal de fala reverberante $s_r(n)$ e da RIR $h(n)$.

Sejam $S_r(k,l)$ e $H(k,l)$ as STFTs de $s_r(n)$ e $h(n)$, respectivamente, utilizando uma janela de Hamming de tamanho M amostras com sobreposição de V amostras, em que $0 \leq l \leq L$ é o índice referente ao segmento temporal, $0 \leq k \leq K$ é o índice referente ao *bin* da DFT, L é o número total de segmentos temporais e K é o número total de *bins* da DFT.

Com base na divisão da RIR em primeiras reflexões e reverberação tardia, $H(k,l)$ é representada por

$$H(k,l) = \begin{cases} B_d(k), & \text{para } l = 0 \\ B_r(k,l)e^{-\tau(k)lR} & \text{para } l > 0, \end{cases} \tag{5.14}$$

em que $B_d(k)$ e $B_r(k,l)$ são variáveis aleatórias Gaussianas de média zero independentes e identicamente distribuídas, $R = M - V$ é o deslocamento entre segmentos consecutivos e $\tau(k)$ é a taxa de decaimento dependente da frequência dada por

$$\tau(k) = \frac{3\ln 10}{T_{60}(k)F_s}, \tag{5.15}$$

em que $T_{60}(k)$ é o tempo de reverberação dependente da frequência. Por definição, $B_d(k)$ contém a informação acerca do caminho direto e poucas primeiras reflexões, e $B_r(k,l)$ contém informação acerca das todas as reflexões posteriores.

Então, a razão de energia direta sobre reverberante dependente da frequência é

$$E_{dr}(k) = 10\log_{10}\left(\frac{1 - e^{-2\tau(k)R}}{e^{-2\tau(k)R}}\frac{1}{\kappa(k)}\right), \tag{5.16}$$

em que $\kappa(k)$ é dado por

$$\kappa(k) = \frac{\mathbf{E}[B_d(k)^2]}{\mathbf{E}[B_r(k,l)^2]}. \tag{5.17}$$

Assumindo que os componentes referentes às primeiras reflexões possuem um número N_e fixo de amostras e expressando a variância espectral dependente da frequência referente a $s_r(n)$ como $\sigma^2_{s_r}(k,l) = \mathbf{E}[|S_r(k,l)|^2]$, conclui-se que a variância espectral da parcela reverberante (primeiras reflexões e reverberação tardia) é

$$\sigma^2_{pt}(k,l) = (1 - \kappa(k))\eta\sigma^2_{pt}(k,l-1) + \kappa(k)\eta\sigma^2_{s_r}(k,l-1), \tag{5.18}$$

em que $\eta = e^{-2\tau(k)R}$ e a variância espectral σ^2_t somente da parcela referente a reverberação tardia é dada por

$$\sigma^2_t(k,l) = e^{-2\tau(k)R(N_e-1)}\sigma^2_{pt}(k,l-N_e+1). \tag{5.19}$$

De maneira similar ao que ocorreu com a estimativa cega em sub-bandas de T_{60} do Capítulo 4, as estimativas de $\sigma^2_t(k,l)$ são convertidas para uma estimativa em banda completa através de um mapeamento. Primeiramente são calculadas as estimativas de cada sub-banda de forma que

$$\sigma^2_t(k) = \sum_{l=0}^{L-1}\sigma^2_t(k,l). \tag{5.20}$$

Em segundo lugar é feita uma primeira estimativa da variância espectral do ambiente. Assim,

$$\bar{\sigma}^2_t = \sum_{k=0}^{K-1}\sigma^2_t(k). \tag{5.21}$$

Por último é calculada a estimativa final de σ_r^2

$$\hat{\sigma}_r^2 = \alpha_\sigma \bar{\sigma}_t^2 + \beta_\sigma, \tag{5.22}$$

em que α_σ e β_σ são obtidos na etapa de treinamento do algoritmo.

Uma desvantagem desse algoritmo de estimação de σ_r^2 é a sua dependência das estimativas de E_{dr} e de T_{60}, uma vez que quaisquer erros de estimação desses parâmetros serão propagados para a estimação de σ_r^2. A nomenclatura associada com essa estimativa depende dos estimadores de T_{60} e E_{dr} utilizados. Foram combinados os 4 estimadores de T_{60} (Ratnam, Vieira, Wen e proposto) com os 2 estimadores de E_{dr} (Falk e proposto) resultando nos seguintes estimadores: $\hat{\sigma}_r^2(R,F)$, $\hat{\sigma}_r^2(V,F)$, $\hat{\sigma}_r^2(W,F)$, $\hat{\sigma}_r^2(P,F)$, $\hat{\sigma}_r^2(R,P)$, $\hat{\sigma}_r^2(V,P)$, $\hat{\sigma}_r^2(W,P)$, $\hat{\sigma}_r^2(P,P)$.

5.4 Estimadores cegos de E_{dr}

5.4.1 Estimador de Falk para E_{dr}

Falk e Chan [56] propõem um estimador cego para E_{dr} que tem como base a representação espectro-temporal conhecida como modulação espectral, que mostra o conteúdo na frequência da envoltória temporal de tempo longo. na

Inicialmente, o sinal de fala reverberante $s_r(n)$ é filtrado por um banco de filtros Gammatone [36] com 23 canais, gerando $s_r^{(j)}(n)$, para $j = 1, 2, \ldots, 23$. Esse banco de filtros tem o objetivo de simular o processamento feito na cóclea. As frequências centrais desse banco de filtros vão de 125 Hz até aproximadamente $\dfrac{F_s}{2}$ e a largura de banda para cada filtro é caracterizada pela banda retangular equivalente.

As saídas $s_r^{(j)}(n)$ do banco de filtros são submetidas à transformada de Hilbert $\mathcal{H}\{\bullet\}$, com o intuito de obter as envoltórias temporais $x^{(j)}(n)$ tais que

$$x^{(j)}(n) = \sqrt{s_r^{(j)}(n)^2 + \mathcal{H}\{s_r^{(j)}(n)\}^2}. \tag{5.23}$$

A envoltória temporal $x_j(n)$ é submetida a uma janela de Hamming de tamanho $M = 0{,}256F_s$ amostras e sobreposição $V = 0{,}875M$, resultando em L segmentos. A DFT dos segmentos é calculada, cujo módulo é chamado $X_j(k,l)$, em que l é o índice referente ao segmento e k é o índice referente ao *bin* da DFT.

A fim de emular o banco de filtros de 8 canais descrito em [60], com as frequências centrais apresentadas na Tabela 5.1, os *bins* de $X_j(k,l)$ são reunidos em 8 grupos, emulando as 8 bandas do filtro de modulação.

Tabela 5.1: Índice i referente à i-ésima banda do filtro de modulação, respectivas frequências centrais e larguras de banda

i	1	2	3	4	5	6	7	8
Frequência central [Hz]	4,0	6,5	10,7	17,6	28,9	47,5	78,1	128,0
Largura de Banda [Hz]	2,4	3,9	6,5	11,0	18,2	29,1	47,6	78,8

Com isto, define-se a energia $\mathcal{X}_{i,j}(l)$ da i-ésima banda de modulação como

$$\mathcal{X}_{i,j}(l) = \sum_{k=K_{m0}(i)}^{K_m(i)} (X_j(k,l))^2, \tag{5.24}$$

em que $K_{m0}(i)$ e $K_m(i)$ são os *bins* inicial e final da DFT referentes à i-ésima banda de modulação. É calculada a média da energia $\mathcal{X}_{i,j}(l)$ dos segmentos com atividade de voz, isto é,

$$\bar{\mathcal{X}}_{i,j} = \frac{1}{L_{\text{act}}} \sum_{l=1}^{L_{\text{act}}} \mathcal{X}_{i,j}^{\text{act}}(l). \tag{5.25}$$

A métrica ORSMR (*overall reverberation-to-speech modulation energy ratio*) é definida como

$$\text{ORSMR} = \frac{\sum_{i=5}^{8} \sum_{j=1}^{23} \bar{\mathcal{X}}^{(i,j)}}{\sum_{j=1}^{23} \bar{\mathcal{X}}^{(1,j)}}. \tag{5.26}$$

Por fim, de acordo com [56] a métrica ORSMR pode ser mapeada em uma estimativa cega para E_{dr} através de

$$\hat{E}_{dr}^{(F)} = 10^{-0,56467-1,644 \log \text{ORSMR}}. \tag{5.27}$$

5.4.2 Estimador proposto para E_{dr}

Esse algoritmo tem como base a estimação da razão de energia direta sobre reverberante dependente da frequência.

Inicialmente estimam-se as FDRs de $s_r(n)$ utilizando uma janela retangular de tamanho M e sobreposição V, gerando um total de L segmentos. Caso ocorra uma região com pelo menos $t_{hr} = \dfrac{0,5F_s}{M}$ segmentos consecutivos com energia decrescente, aquela região é dita uma FDR. Caso nenhuma FDR tenha sido encontrada, t_{hr} é decrescido de um segmento e o processo é repetido até que pelo menos uma região seja encontrada ou $t_{hr} < 3$. Esse procedimento é similar ao adotado para o estimador cego de T_{60} descrito na Seção 4.2. A quantidade de FDRs encontradas no tempo é chamada R_1.

Em seguida é feita a STFT $S_r(k,l)$ de $s_r(n)$, utilizando os mesmos valores de M e V da etapa anterior, limitando os *bins* de modo que a maior frequência analógica utilizada seja 4 kHz. Calcula-se a estimativa $\hat{E}_{dr}(r,k)$ da razão de energia direta sobre reverberante da r-ésima FDR encontrada no tempo e do k-ésimo *bin* da DFT de maneira similar a (2.5).

Outro grupo de FDRs é buscado no domínio da STFT, de maneira similar à feita na Seção 4.2, resultando em $R_2(k)$ FDRs para o k-ésimo *bin* da DFT. Novamente são feitas estimativas $\hat{E}_{dr}(r,k)$, de modo que haverá $R_1 + R_2(k)$ estimativas para cada *bin*.

Então, a estimativa $\hat{E}_{dr}(k)$ do k-ésimo *bin* da DFT é dada por

$$\hat{E}_{dr}(k) = \frac{\sum_{r=1}^{R_1+R_2(k)} \hat{E}_{dr}(r,k)}{R_1 + R_2(k)}. \tag{5.28}$$

A estimativa parcial $\bar{E}_{dr}^{(P)}$ é calculada a partir da média de $\hat{E}_{dr}(k)$, para $\frac{K}{8}+1 \leq k \leq \frac{3K}{8}$, o que equivale a utilizar somente os *bins* da DFT relativos às frequências analógicas no intervalo entre 500 Hz e 1500 Hz. Esse intervalo foi obtido de maneira experimental através do coeficiente de correlação entre as estimativas parciais e as estimativas não cegas $E_{dr}^{(QA)}$, obtidas pelo sistema QAreverb do Capítulo 3 para a base NBP, como mostra a Tabela 5.2.

Tabela 5.2: Coeficiente de correlação ρ entre $E_{dr}^{(QA)}$ e $\bar{E}_{dr}^{(P)}$ utilizando 8 intervalos diferentes

Intervalo	ρ [%]
$0 \leq k \leq K/8$	84
$K/8+1 \leq k \leq 2K/8$	89
$2K/8+1 \leq k \leq 3K/8$	89
$3K/8+1 \leq k \leq 4K/8$	88
$4K/8+1 \leq k \leq 5K/8$	88
$5K/8+1 \leq k \leq 6K/8$	01
$6K/8+1 \leq k \leq 7K/8$	04
$7K/8+1 \leq k \leq 8K/8$	04

A estimativa final $\hat{E}_{dr}^{(P)}$ é dada por

$$\hat{E}_{dr}^{(P)} = \alpha_P \bar{E}_{dr}^{(P)} + \beta_P, \tag{5.29}$$

em que α_p e β_p são constantes obtidas na etapa de treinamento do algoritmo.

5.5 Resultados experimentais

5.5.1 Validação dos estimadores cegos

Essa validação consiste em utilizar a base NBP para analisar de maneira quantitativa a capacidade de estimação dos algoritmos cegos descritos anteriormente.

A Tabela 5.3 mostra o coeficiente de correlação ρ entre as estimativas cegas e não cegas de T_{60}, σ_r^2 e E_{dr}. Além das 8 configurações descritas anteriormente para o estimador de σ_r^2, a configuração $\hat{\sigma}_r^2(Q,Q)$ foi adicionada, significando que foram utilizados os algoritmos não cegos do Capítulo 3, a fim de validar o procedimento sem levar em consideração os erros propagados dos algoritmos utilizados para estimar T_{60} e E_{dr}.

Analisando a Tabela 5.3 nota-se que os estimadores propostos $\hat{T}_{60}^{(P)}$ e $\hat{E}_{dr}^{(P)}$ apresentaram maior capacidade de estimar os parâmetros em questão, o que indica serem as melhores opções para compor a versão cega de Q. Entretanto, todas as configurações para o estimador cego de σ_r^2 apresentaram baixa capacidade para estimar o parâmetro em questão.

Tabela 5.3: Coeficiente de correlação entre os estimadores cegos e não cegos de T_{60}, σ_r^2 e E_{dr}, para a base NBP

Estimador	ρ [%]
$\hat{T}_{60}^{(R)}$	55
$\hat{T}_{60}^{(V)}$	64
$\hat{T}_{60}^{(W)}$	12
$\hat{T}_{60}^{(P)}$	91
$\hat{\sigma}_r^2(Q,Q)$	63
$\hat{\sigma}_r^2(R,F)$	44
$\hat{\sigma}_r^2(R,P)$	42
$\hat{\sigma}_r^2(V,F)$	27
$\hat{\sigma}_r^2(V,P)$	42
$\hat{\sigma}_r^2(W,F)$	4
$\hat{\sigma}_r^2(W,P)$	3
$\hat{\sigma}_r^2(P,F)$	41
$\hat{\sigma}_r^2(P,P)$	47
$\hat{E}_{dr}^{(F)}$	17
$\hat{E}_{dr}^{(P)}$	89

A Tabela 5.4 apresenta o coeficiente de correlação entre a versão cega $Q_b(A1,A2,A3)$ de (3.1) e a nota MOS para a base NBP, em que $A1 = R,V,W,P$ representa as 4 possibilidades de estimadores de T_{60} (Ratnam, Vieira, Wen e proposto), $A2 = RF,RP,VF,VP,WF,WP,PF,PP$ representa as 8 possíveis configurações para o estimador de σ_r^2 e $A3 = F,P$ representa as 2 possibilidades de estimadores de E_{dr} (Falk e proposto). É possível notar que o melhor resultado se dá

para as configurações $Q_b(P, PP, P)$ e $Q_b(P, PF, F)$, ambos utilizando o estimador de T_{60} proposto nesta tese. Como a correlação é numericamente igual para essas duas configurações, a base MARDY foi utilizada a fim de descobrir se as configurações são equivalentes ou se há uma indicação de superioridade de desempenho entre elas.

Tabela 5.4: Coeficiente de correlação entre $Q_b(A1, A2, A3)$ e a nota MOS para a base NBP.

$Q_b(A1, A2, A3)$	ρ [%]
$Q_b(R, RF, F)$	59
$Q_b(R, RP, P)$	72
$Q_b(V, VF, F)$	57
$Q_b(V, VP, P)$	63
$Q_b(W, WF, F)$	5
$Q_b(W, WP, P)$	5
$Q_b(P, PF, F)$	87
$Q_b(P, PP, P)$	87

A Tabela 5.5 mostra o coeficiente de correlação entre as configurações $Q_b(P, PF, F)$ e $Q_b(P, PP, P)$ e a nota MOS, para a base MARDY. A partir dessa tabela, fica claro que a configuração $Q_b(P, PP, P)$ possui o melhor desempenho dentre todas as configurações testadas.

Tabela 5.5: Coeficiente de correlação entre algumas configurações de $Q_b(A1, A2, A3)$ e a nota MOS, para a base MARDY

$Q_b(A1, A2, A3)$	ρ [%]
$Q_b(P, PF, F)$	83
$Q_b(P, PP, P)$	91

Os resultados mostrados nas Tabelas 5.3, 5.4 e 5.5 confirmam que a melhor configuração para a versão cega de (3.1) é $Q_b(P, PP, P)$, então, a versão cega de (3.1) é dada por

$$Q_b = \frac{\hat{T}_{60}^{(P)} \hat{\sigma}_r^2(P, P)}{(\hat{E}_{dr}^{(P)})^{0,3}}. \tag{5.30}$$

Assim como na versão não cega, a métrica Q_b é mapeada em Q_b^{MOS} através de um mapeamento de terceira ordem, para que a métrica final fique na escala MOS, de modo que

$$Q_b^{\text{MOS}} = \alpha_b(x_{b1}Q_b^3 + x_{b2}Q_b^2 + x_{b3}Q_b + x_{b4}) + \beta_b, \tag{5.31}$$

em que os coeficientes α_b, β_b, x_{b1}, x_{b2}, x_{b3} e x_{b4} são determinados durante o treinamento do sistema.

5.5.2 Comparação com outras métricas de avaliação de qualidade

A Tabela 5.6 mostra o coeficiente de correlação entre as notas subjetivas MOS e algumas métricas objetivas, para as bases NBP e MARDY. Essas métricas objetivas são divididas em não cegas (W-PESQ, R_{DT} e Q_{MOS}) e cegas (P.563, SRMR e Q_b^{MOS}). Todas as métricas foram avaliadas em suas versões com e sem mapeamento de terceira ordem, feito de maneira independente para cada métrica utilizando a metade dos sinais da base NBP.

A partir da Tabela 5.6, conclui-se que a métrica cega proposta nas suas versões sem mapeamento Q_b e com mapeamento Q_b^{MOS} tem um melhor desempenho em relação às outras métricas cegas analisadas, indicando uma maior capacidade em estimar a métrica subjetiva MOS. A única métrica analisada que obteve um melhor desempenho foi a métrica não cega na sua versão com mapeamento Q_{MOS}, também proposta nesta tese, o que é esperado, uma vez que Q_b^{MOS} é a versão cega da métrica utilizada pelo sistema QAreverb.

Tabela 5.6: Coeficiente de correlação ρ entre as notas subjetivas MOS e algumas métricas objetivas (com mapeamento/sem mapeamento), para as bases NBP e MARDY

Coeficiente de correlação (ρ [%])		
	Bases	
Métrica	NBP	MARDY
W-PESQ	77/89	72/77
R_{DT}	59/61	64/64
Q / Q_{MOS}	81/91	95/95
P.563	52/59	55/54
SRMR	74/81	79/77
Q_b / Q_b^{MOS}	87/87	91/91

As Figs. 5.1 e 5.2 mostram a comparação entre as notas subjetivas MOS (linha sólida) e a métrica cega proposta Q_b^{MOS} (pontos) para todos os sinais das bases NBP e MARDY, respectivamente. Essas figuras comprovam capacidade da métrica proposta em estimar de maneira cega a qualidade percebida de sinais de fala contaminados com reverberação.

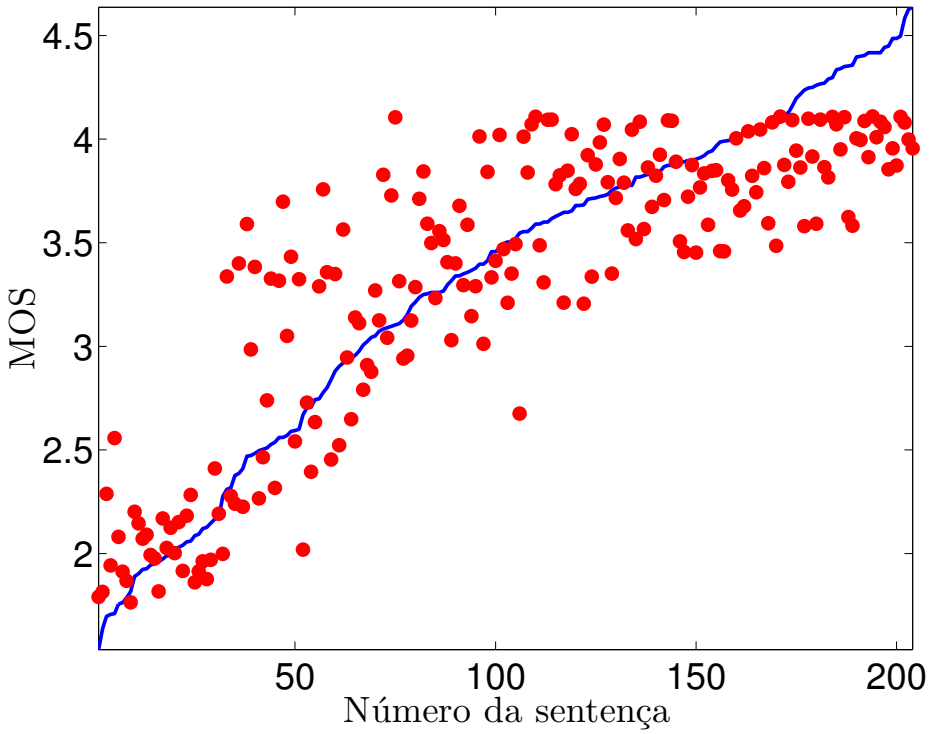


Figura 5.1: Curvas das notas subjetivas MOS (linha sólida) e das notas objetivas Q_b^{MOS} (pontos), de todos os sinais da base NBP.

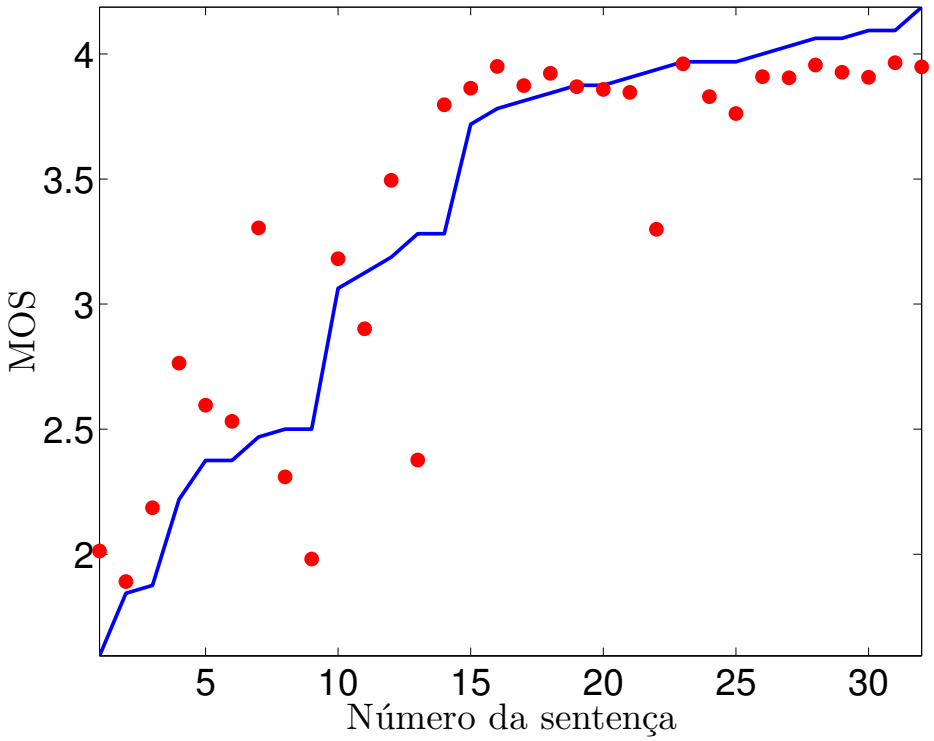


Figura 5.2: Curvas das notas subjetivas MOS (linha sólida) e das notas objetivas Q_b^{MOS} (pontos), de todos os sinais da base MARDY.

5.6 Conclusões

Este capítulo descreveu de maneira detalhada a proposta de uma versão cega, isto é, sem a utilização dos sinais de referência, do algoritmo não cego descrito no Capítulo 3. A métrica proposta tem como base três estimadores cegos $\hat{T}_{60}^{(P)}$, $\hat{\sigma}_r^2(P,P)$ e $\hat{E}_{dr}^{(P)}$ para tempo de reverberação, variância espectral do ambiente e razão de energia direta sobre reverberante, respectivamente.

O algoritmo proposto obteve um coeficiente de correlação entre a métrica Q_b^{MOS} e a nota MOS de 87% e 91% para as bases NBP e MARDY, respectivamente. Esses resultados indicam que a técnica proposta tem maior capacidade de estimar de maneira cega a qualidade percebida de sinais de fala reverberantes, em relação a outros algoritmos cegos encontrados na literatura.

A abordagem proposta pode ser utilizada para monitorar a quantidade de reverberação em sistemas de comunicação e para avaliar ou desenvolver sistemas automáticos para desreverberação de sinais de fala reverberantes.

Capítulo 6

Desreverberação sem referência

6.1 Introdução

A reverberação pode afetar decisivamente o desempenho dos atuais sistemas de reconhecimento de fala/locutor ou mesmo dos sistemas de auxílio à reabilitação de deficientes auditivos. Isso motiva a utilização de técnicas apropriadas para reduzir os seus efeitos. Embora a reverberação possa ser bastante prejudicial, em pequena quantidade/intensidade ela pode até tornar a fala mais agradável para o ouvinte comum [61]. A utilização de arranjo de microfones é a configuração mais comumente usada em técnicas de desreverberação, porém para as aplicações anteriormente mencionadas a utilização de um único microfone é mais apropriada.

Neste capítulo é proposta uma metologia para o aperfeiçoamento de técnicas de desreverberação. Para exemplificar tal metodologia, são sugeridas análises e propostas de modificações para o algoritmo de desreverberação de sinais de fala usando um único microfone como proposto em [3]. Esse algoritmo é dividido em dois blocos: o primeiro lida com os efeitos das primeiras reflexões e o segundo lida com os efeitos da reverberação tardia. A influência das primeiras reflexões é reduzida através de um processo adaptativo de filtragem inversa, que tem o objetivo de reconstruir o sinal de fala desejado. Já o efeito dos componentes de reverberação tardia é mitigado usando-se subtração espectral, onde um modelo baseado na distribuição de Rayleigh é utilizado para emular o comportamento dos componentes de reverberação tardia.

Na Seção 6.2, é apresentada uma visão geral do algoritmo original de desreverberação de dois estágios, além da descrição dos estágios de filtragem inversa e de subtração espectral. A Seção 6.3 descreve as modificações propostas para o estágio de filtragem inversa, assim como os sinais utilizados para o treinamento e teste do sistema. Na Seção 6.4, são descritas as bases utilizadas no desenvolvimento e validação das alterações propostas para o estágio de subtração espectral, as modificações propostas e seus resultados experimentais. A Seção 6.5 descreve a

combinação das modificações no estágio de filtragem adaptativa com as modificações no estágio de subtração espectral e seus resultados experimentais. Finalmente, as conclusões referentes ao desempenho das modificações propostas estão na Seção 6.7.

6.2 Algoritmo de 2 estágios

O algoritmo de dois estágios descrito em [3] consiste na utilização de blocos isolados de processamento de sinais com o intuito de reduzir o nível de reverberação do sinal de saída quando comparado com o sinal reverberante originalmente aplicado à entrada. Os dois blocos do algoritmo são chamados de filtragem inversa e subtração espectral, como mostrado na Fig. 6.1, onde $s_r(n)$, $\tilde{s}(n)$ e $\hat{s}(n)$ são, respectivamente, os sinais de fala reverberante, inversamente filtrado e subtraído espectralmente (chamado neste trabalho de desreverberado).

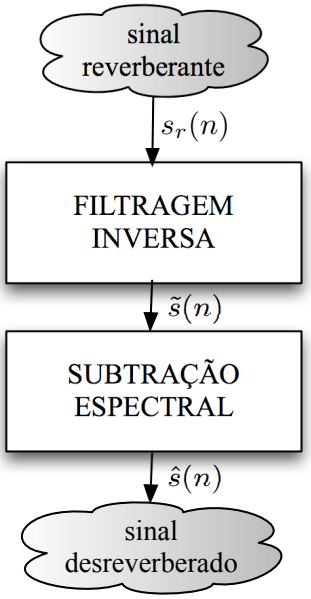


Figura 6.1: Diagrama de blocos do algoritmo de dois estágios.

Como visto na Seção 2.2, o efeito de reverberação ocorre devido às reflexões que o sinal original sofre nas superfícies do ambiente e a RIR pode ser dividida em primeiras reflexões e reverberação tardia. As primeiras reflexões possuem um caráter de impulsos esparsos, geralmente são modeladas utilizando *tapped delay lines* [62], e terminam quando a reverberação atinge um comportamento estatístico assintótico. A reverberação tardia é caracterizada por uma alta densidade de impulsos, decorrentes das reflexões das reflexões do sinal original nas superfícies do ambiente. Então, o algoritmo dado em [3] lida com esses dois tipos de efeitos de reverberação em dois estágios: filtragem inversa, que tem por objetivo mitigar o efeito das primeiras reflexões, e subtração espectral, que tem por objetivo mitigar o efeito da reverberação tardia.

6.2.1 Filtragem inversa

O objetivo da filtragem inversa é reconstruir uma estimativa do sinal original de fala, reduzindo os efeitos da coloração. Essa etapa tem como base o trabalho de Gillespie et al. [63], no qual é proposto um algoritmo que utiliza múltiplos microfones para calcular um filtro inverso através da maximização da curtose do resíduo de predição linear, gerando, assim, um modelo inverso para RIR associada.

Seja $\mathbf{g} = [g_0 \; g_1 \; \ldots \; g_{M-1}]^T$ o filtro inverso de comprimento M cuja resposta ao impulso é dada por

$$h_g(n) = \sum_{m=0}^{M-1} g_m \delta(n-m). \tag{6.1}$$

O sinal de fala filtrado inversamente $\tilde{s}(n)$ é dado por

$$\tilde{s}(n) = h_g(n) * s_r(n). \tag{6.2}$$

Como o resíduo de predição linear do sinal de fala limpo (sem reverberação) possui valor de curtose maior do que o do sinal de fala reverberante, o filtro inverso $\mathbf{g}$ pode ser obtido através da maximização da curtose do resíduo de predição linear $s_{rp}(n)$ do sinal de fala reverberante $s_r(n)$, sendo o resíduo de predição linear de $\tilde{s}(n)$ dado por

$$\tilde{s}_p(n) = h_g(n) * s_{rp}(n). \tag{6.3}$$

Para evitar o aparecimento de artefatos em $\tilde{s}(n)$ devido à reconstrução a partir do resíduo de predição linear, primeiramente determina-se $\mathbf{g}$ a partir de $s_{rp}(n)$ e depois obtém-se $\tilde{s}(n)$ a partir de (6.2).

A otimização do filtro inverso é feita de acordo com um algoritmo similar ao algoritmo adaptativo LMS (*least mean square*) em segmentos de comprimento M. O l-ésimo segmento de $s_{rp}(n)$, definido por $\mathbf{s_{rp}}(l) = [s_{rp}(lM) \; \ldots \; s_{rp}((l+1)M-1)]^T$, e o l-ésimo segmento de $\tilde{s}_p(n)$ é dado por $\tilde{\mathbf{s}}_\mathbf{p}(l) = [\tilde{s}_p(lM) \; \ldots \; \tilde{s}_p((l+1)M-1)]^T$.

Seja $J(l)$ a curtose de $\tilde{\mathbf{s}}_\mathbf{p}(l)$ dada por

$$J(l) = \frac{\mathbf{E}[\tilde{\mathbf{s}}_\mathbf{p}^4(l)]}{\mathbf{E}^2[\tilde{\mathbf{s}}_\mathbf{p}^2(l)]} - 3, \tag{6.4}$$

em que $\tilde{\mathbf{s}}_\mathbf{p}^k(l) = [\tilde{s}_p^k(lM) \; \ldots \; \tilde{s}_p^k((l+1)M-1)]^T$. O algoritmo adaptativo utiliza a média de $J(l)$ como função custo e é dada por

$$\bar{J} = \frac{1}{L}\sum_{l=0}^{L-1} J(l) = \frac{1}{L}\sum_{l=0}^{L-1}\left(\frac{\mathbf{E}[\tilde{\mathbf{s}}_\mathbf{p}^4(l)]}{\mathbf{E}^2[\tilde{\mathbf{s}}_\mathbf{p}^2(l)]} - 3\right), \tag{6.5}$$

em que L é o número total de segmentos.

Seja $\mathbf{f}(l)$ o gradiente de $J(l)$, temos que

$$\begin{aligned}\mathbf{f}(l) &= \nabla J(l) \\ &= \frac{4(\mathbf{E}[\tilde{\mathbf{s}}_{\mathbf{p}}^2(l)]\tilde{\mathbf{s}}_{\mathbf{p}}^3(l) - \mathbf{E}[\tilde{\mathbf{s}}_{\mathbf{p}}^4(l)]\tilde{\mathbf{s}}_{\mathbf{p}}(l))}{\mathbf{E}^3[\tilde{\mathbf{s}}_{\mathbf{p}}^2(l)]}.\end{aligned} \tag{6.6}$$

De acordo com Haykin [64], a implementação desse tipo de formulação no domínio do tempo não é recomendada, uma vez que um grande espalhamento dos autovalores da matriz de correlação do sinal de entrada pode causar problemas de convergência. Por esse motivo, é empregada uma estrutura no domínio da frequência, em que $\mathbf{G}(\mathbf{i})$, $\mathbf{F}(l)$ e $\mathbf{Y_r}(l)$ são as DFTs da i-ésima iteração $\mathbf{g}^{(i)}$, de $\mathbf{f}(l)$ e de $\mathbf{s_{rp}}(l)$, respectivamente.

Logo, a equação de adaptação do filtro inverso é dada por

$$\mathbf{G}(i+1) = \mathbf{G}(i) + \frac{\mu}{L}\sum_{l=0}^{L-1}\mathbf{F}(l)\mathbf{Y}_{\mathbf{r}}^*(l), \tag{6.7}$$

em que μ é o passo de adaptação e o asterisco sobrescrito denota o complexo conjugado. Considerando aspectos práticos, o algoritmo descrito em [3] utiliza passo de adaptação $\mu = 3 \times 10^{-9}$, comprimento do filtro de predição linear $R = 10$ e comprimento do segmento de análise $M = 0{,}032F_s$ amostras com sobreposição $V = 0{,}5M$.

6.2.2 Subtração espectral

A Fig. 6.2 detalha os componentes do bloco de subtração espectral, onde se considera que as reflexões inicias e tardias são aproximadamente descorrelacionadas [3]. Esse bloco utiliza como entrada o sinal de fala inversamente filtrado $\tilde{s}(n)$ e tem como saída o sinal de fala desreverberado $\hat{s}(n)$. É importante ressaltar que a fase do sinal inversamente filtrado é usada para gerar o sinal desreverberado.

A subtração espectral tem por objetivo reduzir o efeito da reverberação de longa duração causado pelo componente de reverberação tardia da RIR. Inicialmente o sinal inversamente filtrado $\tilde{s}(n)$ é particionado em segmentos usando uma janela de Hamming de 32 ms com 24 ms de sobreposição entre segmentos consecutivos.

Seja $S_{\tilde{s}}(k,l) = |S_{\tilde{s}}(k,l)|e^{j\varphi_{\tilde{s}}(k,l)}$ a STFT do sinal de fala inversamente filtrado $\tilde{s}(n)$, em que $k \in \mathbb{N}$ é o índice do *bin* da STFT e $l \in \mathbb{N}$ é o índice do segmento. Seja ainda $w(l)$ uma janela de atenuação cuja função tem a forma da distribuição de Rayleigh, dada por

$$\begin{cases} w(l) = \left(\frac{l+a}{a^2}\right) e^{\left(\frac{-(l+a)^2}{2a^2}\right)}, & \text{se } l > -a \\ w(l) = 0, & \text{caso contrário,} \end{cases} \tag{6.8}$$

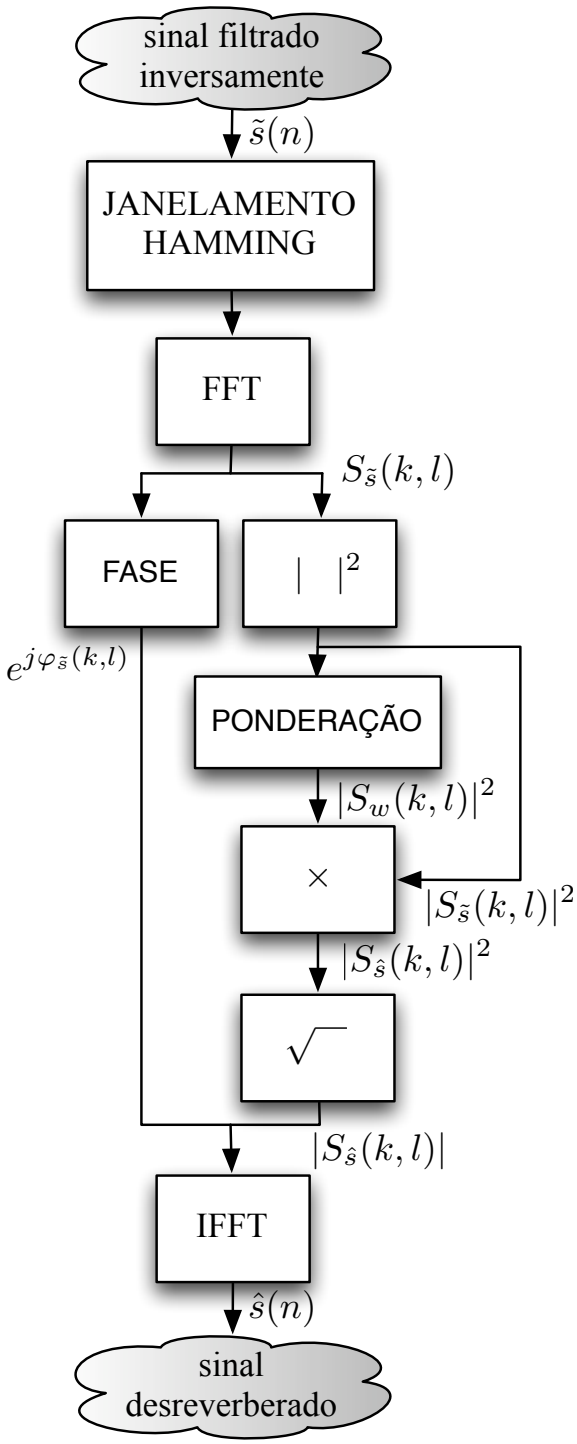


Figura 6.2: Diagrama de blocos da etapa de subtração espectral.

em que a controla o espalhamento total da função.

Se ϕ é o comprimento das primeiras reflexões e ζ o fator de escala que estabelece a energia relativa do componente de reverberação tardia depois da filtragem inversa, o espectro de potência da reverberação tardia é modelado por

$$|S_l(k,l)|^2 = \sum_{\tau=-\infty}^{\infty} \zeta w(\tau - \phi)|S_{\tilde{s}}(k, l-\tau)|^2. \tag{6.9}$$

Esse modelo, dado em (6.9), foi baseado no efeito de distorção dos componentes de reverberação tardia, que causam atenuação no espectro de sinal. O que resulta em o espectro de potência das componentes tardias ser uma versão atenuada e deslocada no tempo do espectro de potência do sinal de fala inversamente filtrado.

Considerando que os componentes das primeiras reflexões e da reverberação tardia são descorrelacionados, o espectro de potência dos primeiros componentes pode ser estimado pela subtração do espectro de potência dos componentes de reverberação tardia do sinal inversamente filtrado. O módulo de subtração espectral faz uma espécie de ponderação no espectro de potência de $\tilde{s}(n)$, onde o bloco de PONDERAÇÃO da Fig. 6.2 é dado por

$$|S_w(k,l)|^2 = \max\left[1 - \frac{|S_l(k,l)|^2}{|S_{\tilde{s}}(k,l)|^2}, \xi\right], \tag{6.10}$$

$\xi = 0{,}001$ correspondendo à máxima atenuação de 30 dB e finalmente o espectro de potência de $\hat{s}(n)$ sendo

$$|S_{\hat{s}}(k,l)|^2 = |S_{\tilde{s}}(k,l)|^2 \times |S_w(k,l)|^2. \tag{6.11}$$

No intuito de se calcular $\hat{s}(n)$, a informação de fase obtida de $S_{\tilde{s}}(k,l)$ é combinada com o módulo de $S_{\hat{s}}(k,l)$. Assim,

$$S_{\hat{s}}(k,l) = |S_{\hat{s}}(k,l)|e^{j\varphi_{\tilde{s}}(k,l)} \tag{6.12}$$

Os valores originais das constantes utilizadas em [3] são $\phi = 7$, $a = 5$, $\zeta = 0{,}35$ e $\xi = 0{,}001$.

6.3 Modificações na filtragem inversa

Nesta seção, são considerados três aspectos do estágio de filtragem inversa:

1. A ordem R do filtro de predição linear.
2. A influência do passo de adaptação μ.
3. O critério de convergência

Todas as modificações propostas neste capítulo utilizam uma sub-base (intitulada base de treinamento) da NBP contendo 18 sinais, um para cada ambiente:

- Anecoico
- 6 RIRs artificiais

- 4 salas naturais
- 7 salas reais

O algoritmo original mantém a qualidade estimada $Q_{\text{MOS}} = 3{,}46$ em relação à base de treinamento não processada.

6.3.1 Influência da ordem R do filtro de predição linear

A análise da ordem R do filtro de predição linear foi motivada pela observação do comportamento da curtose média $\bar{J}(i)$ do resíduo $\tilde{\mathbf{s}}_{\mathbf{p}}(l)$ da i-ésima iteração. Esse comportamento é ilustrado na Fig. 6.3, que mostra o valor de $\bar{J}(i)$ para um sinal da NBP com $T_{60} = 920$ ms. Nessa figura, observa-se que, após a convergência, o valor de $\bar{J}(i)$ oscila ao redor de um valor médio. A amplitude da oscilação é tal que valores distintos de $\bar{J}(i)$ acarretam desempenhos distintos, o que é indesejável.

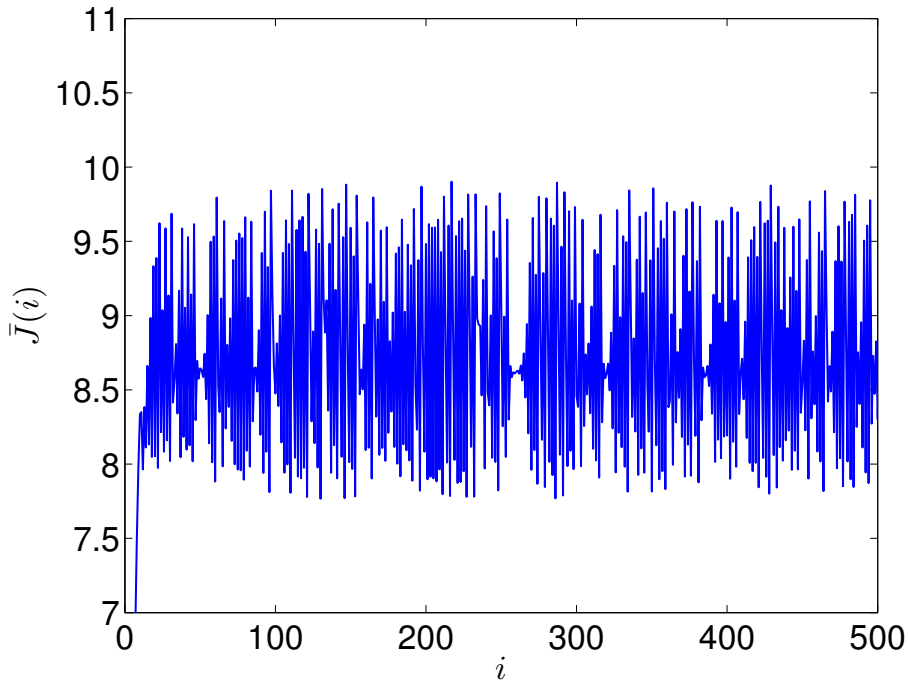


Figura 6.3: Convergência da curtose média $\bar{J}(i)$ utilizando $R = 10$ coeficientes de predição linear, $\mu = 3 \times 10^{-9}$ e um total de $N_i = 500$ iterações.

O aumento da ordem do filtro de predição linear tende a resultar em um resíduo mais impulsivo, acompanhando os ciclos glotais, sendo essa uma estrutura mais apropriada a ser otimizada pela maximização da curtose.

A Fig. 6.4 mostra o resultado médio da métrica Q_{MOS} para os sinais desreverberados da base de treinamento, variando-se a ordem do filtro de predição linear de $R = 10$ a $R = 100$, com os valores originais $\mu = 3 \times 10^{-9}$ e $N_i = 500$ iterações. A

partir dessa figura, observa-se que o melhor desempenho se deu para $R = 30$ coeficientes, correspondendo a um perfil de curtose mais suave, como mostrado na Fig. 6.5.

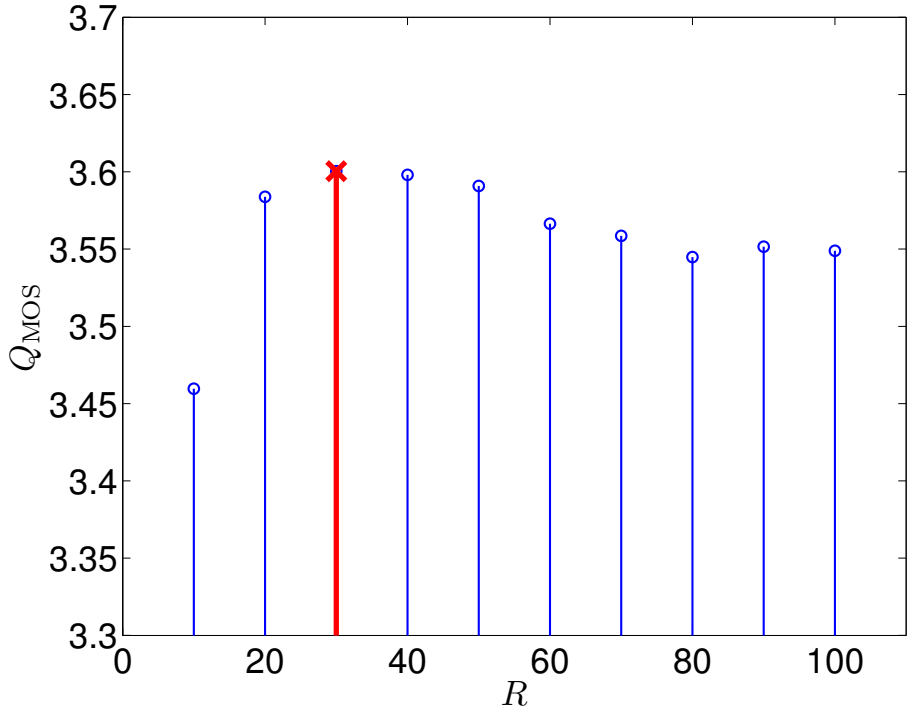


Figura 6.4: Qualidade do sinal desreverberado avaliado a partir da métrica Q_{MOS} como função da ordem R do filtro de predição linear com $\mu = 3 \times 10^{-9}$ e um total de $N_i = 500$ iterações. O melhor desempenho se deu para $R = 30$ coeficientes, marcado com um $\times$ no gráfico.

6.3.2 Influência do passo de adaptação μ

A Fig. 6.6 ilustra a métrica Q_{MOS} para diversos valores de μ dentro do intervalo $[1 \times 10^{-9}; 10 \times 10^{-9}]$, para $R = 30$ e $N_i = 500$. Nota-se que o melhor desempenho se dá para o passo de adaptação $\mu = 1 \times 10^{-9}$, que resulta em um perfil ainda mais suave para $\bar{J}(i)$, como mostra a figura 6.7

6.3.3 Critério de convergência para o algoritmo adaptativo

A fim de evitar um custo computacional desnecessário ao se fixar o número de iterações em $N_i = 500$, um novo critério de parada foi desenvolvido para esse algoritmo, com o objetivo de diminuir o número médio de iterações necessárias para realizar a filtragem inversa, mantendo qualidade similar ao caso com o número de iterações fixo.

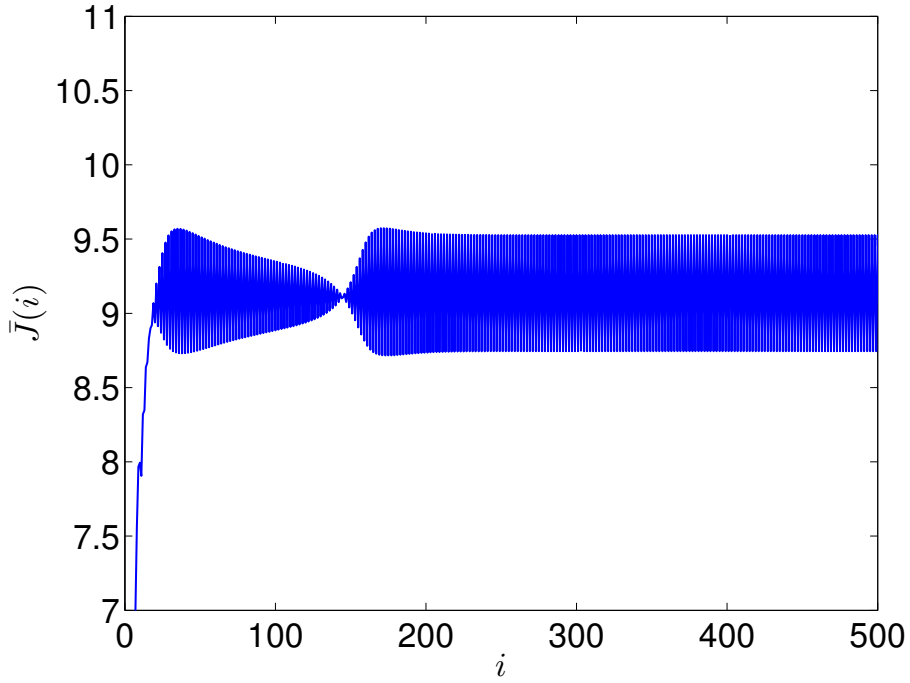


Figura 6.5: Convergência da curtose média $\bar{J}(i)$ utilizando $R = 30$ coeficientes de predição linear, $\mu = 3 \times 10^{-9}$ e um total de $N_i = 500$ iterações.

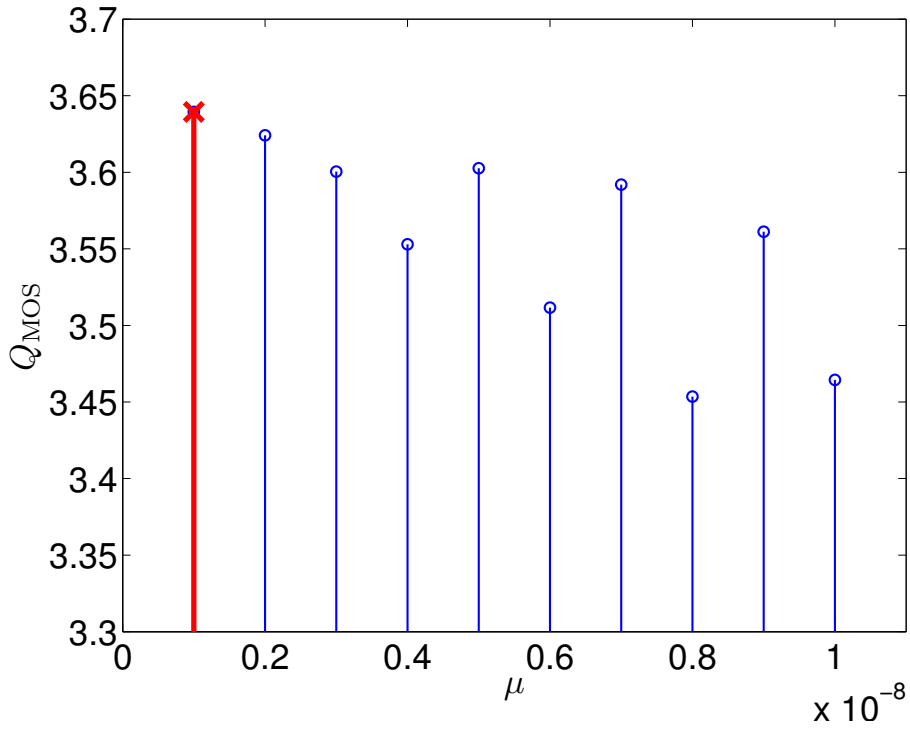


Figura 6.6: Qualidade do sinal desreverberado avaliado a partir da métrica Q_{MOS} como função do passo de adaptação μ com $R = 30$ coeficientes de predição linear e um total de $N_i = 500$ iterações.

Seja a variação no tempo da curtose média dada por

$$\bar{J}_d(i) = \frac{\left|\sum_{l=1}^{\bar{L}} \bar{J}(i-l) - \sum_{l=1}^{\bar{L}} \bar{J}(i-l+1)\right|}{\left|\sum_{l=1}^{\bar{L}} \bar{J}(i-l)\right|}. \tag{6.13}$$

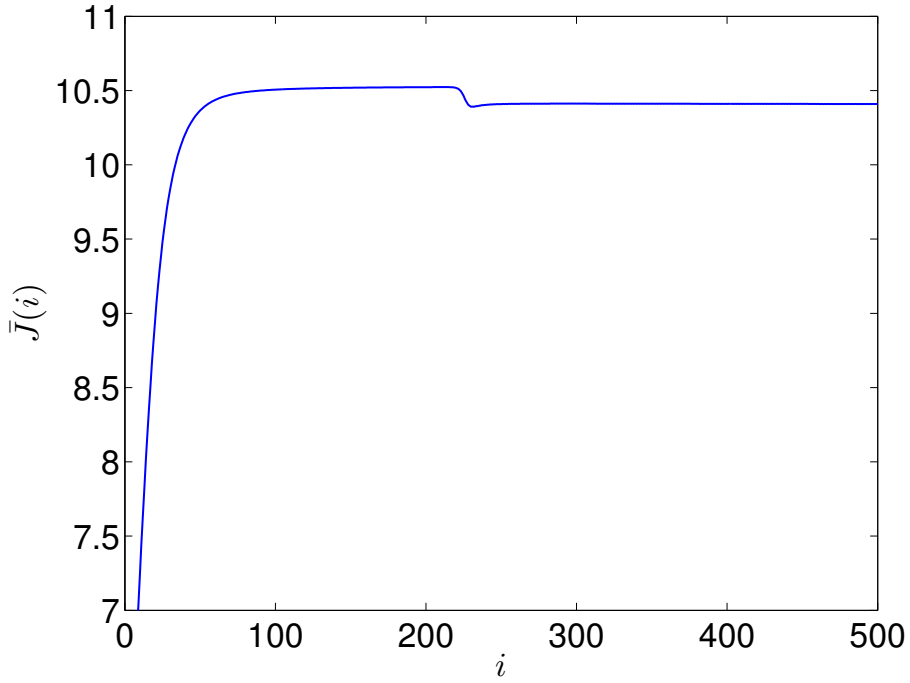


Figura 6.7: Convergência da curtose média $\bar{J}(i)$ utilizando $R = 30$ coeficientes de predição linear, $\mu = 1 \times 10^{-9}$ e um total de $N_i = 500$ iterações.

O algoritmo é parado quando $\bar{J}_d(i)$ atinge um certo limiar J_{thr}, com $\bar{L} = 4$, obtidos experimentalmente para garantir um processo suave.

A variação da curtose para $R = 30$ e $\mu = 1 \times 10^{-9}$ é mostrada na Fig. 6.8, em que se nota um padrão decrescente até aproximadamente -100 dB. Para a base de treinamento, o valor de $J_{thr} = -25$ dB resulta em uma média de $N_i = 5{,}4$ iterações e em uma média de $Q_{\text{MOS}} = 3{,}68$, representando queda de aproximadamente 99% da complexidade computacional e um aumento de Q_{MOS} em aproximadamente 6% em relação ao algoritmo original.

6.3.4 Combinação

Para escolher uma melhor configuração para a filtragem inversa, fez-se uma busca exaustiva dentro dos seguintes intervalos:

$$\begin{cases} R &= \{10, 20, \ldots, 100\} \\ \mu &= \{1, 2, \ldots, 10\} \times 10^{-9} \\ J_{thr} &= \{-\infty, -100, -50, -25\} \text{ dB} \end{cases} .$$

A configuração com o maior valor para a média de $Q_{\text{MOS}} = 3{,}71$ e com a média de $N_i \leq 50$ iterações foi $R = 40$, $\mu = 4 \times 10^{-9}$, $J_{thr} = -25$ dB.

Foram selecionadas as 19 configurações com o maior valor médio de Q_{MOS}, além da configuração original, (intituladas 19+1 configurações da filtragem inversa) para

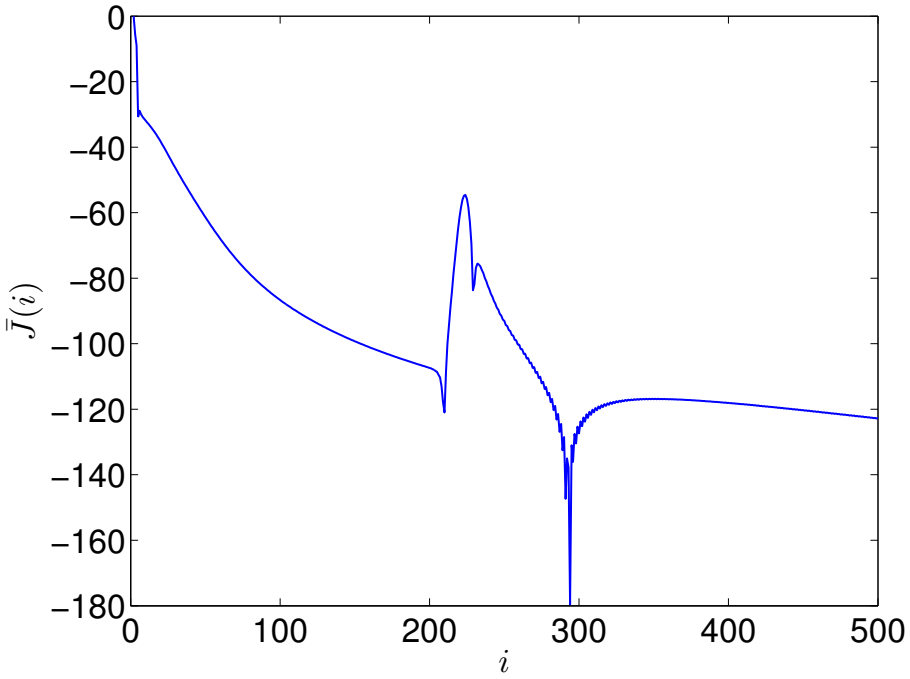


Figura 6.8: Variação da curtose média com $\bar{L} = 4$, $R = 30$ coeficientes de predição linear e $\mu = 1 \times 10^{-9}$.

serem combinadas as 19+1 configurações da subtração espectral (explicadas na Seção 6.4.3).

6.3.5 Validação

Para validar a configuração escolhida, aplicou-se o algoritmo de dois estágios com a configuração original $\{R = 10,\ \mu = 3 \times 10^{-9}$ e $J_{thr} = -\infty$ dB$\}$ e FI (filtragem inversa modificada) $\{R = 40,\ \mu = 4 \times 10^{-9}$ e $J_{thr} = -25$ dB$\}$ nos 200 sinais não anecoicos da base NBP.

A Tabela 6.1 mostra as médias de Q_{MOS}, T_{60}, σ_r^2 e E_{dr} estimados através do sistema QAreverb descrito no Capítulo 3, para as configurações original e FI.

Tabela 6.1: Valores médios de Q_{MOS}, T_{60}, σ_r^2 e E_{dr} para as versões original e modificada do algoritmo de dois estágios

Métrica utilizada	Sinais não processados	Configuração	
		Original	FI
Q_{MOS}	3,36	3,42	3,52
T_{60} [ms]	517	337	368
σ_r^2	5,61	6,80	6,05
E_{dr}	7,59	2,33	4,61

Comparando os desempenhos para os 200 sinais, nota-se que há melhora de 3%, 11% e 98%, porém há piora de 9% para as médias de Q_{MOS}, σ_r^2, E_{dr} e T_{60},

respectivamente. Comparando os resultados obtidos para a versão modificada com os parâmetros da base não processada, percebe-se que há melhora de 5% e 29%, porém há piora de 8% e 61% para as médias de Q_{MOS}, T_{60}, σ_r^2 e E_{dr}, respectivamente.

A média do número de iterações N_i para os 200 sinais foi de 5,50 iterações para a versão modificada, representando uma redução média na complexidade computacional de aproximadamente 99%.

6.4 Modificações na subtração espectral

Esta seção analisa as modificações do algoritmo, dado em [3], no bloco de subtração espectral da seguinte forma:

1. São feitas duas otimizações conjuntas: fator de atenuação ζ com limiar ξ; e atraso ϕ com fator de espalhamento a.

2. São selecionadas 15 configurações para cada uma das otimizações anteriores.

3. As 15 configurações de cada otimização em par são combinadas em 225 possíveis configurações para a subtração espectral.

Todas as propostas de modificações estão associadas a (6.10), sendo que a última está também relacionada com a distribuição de Rayleigh como mostrado em (6.8).

A otimização do fator de atenuação ζ pode proporcionar uma melhor relação entre a energia dos primeiros componentes e dos componentes da reverberação tardia, proporcionando uma maior redução perceptual dos efeitos desta última.

A otimização do limiar ξ deve afetar a representação espectral do sinal desreverberado, fazendo com que uma representação não apropriada influencie fortemente na qualidade perceptual do sinal processado.

A otimização de ϕ e a, que são o tamanho em segmentos das primeiras reflexões e o fator de espalhamento da distribuição de Rayleigh, respectivamente, também afeta (6.10). Os valores de a são dependentes de ϕ, uma vez que a não pode ser maior do que ϕ. A escolha apropriada desses parâmetros vai estabelecer o instante de tempo adequado para a aplicação da atenuação dos componentes da reverberação tardia e o formato da janela de atenuação desses componentes.

6.4.1 Otimização conjunta do fator de atenuação ζ e do limiar ξ

Inicialmente foi realizada uma busca conjunta pelos valores de ζ e ξ, para depois serem selecionadas os 14 pares $[\zeta, \xi]$ com os maiores valores médios de Q_{MOS}, além do par original, intitulados 14+1 pares $[\zeta, \xi]$ para serem utilizados na otimização conjunta de ϕ, a, ζ e ξ.

6.4.2 Otimização conjunta do atraso ϕ e do fator de espalhamento a

Inicialmente foi realizada uma busca conjunta pelos valores de ϕ e a, para depois serem selecionadas os 14 pares $[\phi, a]$ com os maiores valores médios de Q_{MOS}, além do par original, intitulados 14+1 pares $[\phi, a]$ para serem utilizados na otimização conjunta de ϕ, a, ζ e ξ.

6.4.3 Otimização conjunta de ϕ, a, ζ e ξ

A otimização conjunta de ϕ, a, ζ e ξ resultou em uma média de $Q_{\text{MOS}} = 3{,}61$, para a configuração $\phi = 7$, $a = 6$, $\zeta = 0{,}35$ e $\xi = 0.001$.

Os 14+1 pares $[\phi, a]$ foram combinados com os 14+1 pares $[\zeta, \xi]$, resultando em um total de 225 possíveis combinações, das quais as 19 com o maior valor médio de Q_{MOS}, além da configuração original, (intituladas 19+1 configurações da subtração espectral) foram selecionadas para serem combinadas com as 19+1 configurações da filtragem inversa.

6.4.4 Validação

Nesta seção, a análise é realizada usando-se a base de teste para avaliar a qualidade percebida das versões original $\{\zeta = 0{,}35,\ \xi = 0{,}001,\ \phi = 7$ e $a = 5\}$ e SE (subtração espectral modificada) $\{\zeta = 0{,}35,\ \xi = 0{,}001,\ \phi = 7$ e $a = 6\}$ do algoritmo de desreverberação de dois estágios. O objetivo é utilizar os 200 sinais não anecoicos da NBP para validar os valores ótimos dos parâmetros obtidos utilizando a base de treinamento.

A Tabela 6.2 mostra as médias de Q_{MOS}, T_{60}, σ_r^2 e E_{dr} estimadas através do sistema QAreverb descrito no Capítulo 3, para as configurações original e SE.

Tabela 6.2: Valores médios de Q_{MOS}, T_{60}, σ_r^2 e E_{dr} para as versões original e modificada do algoritmo de dois estágios

Métrica utilizada	Sinais não processados	Configuração	
		Original	SE
Q_{MOS}	3,36	3,42	3,43
T_{60} [ms]	517	337	340
σ_r^2	5,61	6,80	6,76
E_{dr}	7,59	2,33	2,38

Comparando os desempenhos para os 200 sinais, nota-se que há melhora de 0,3%, 0,9% e 2%, porém há piora de 0,5% para as médias de Q_{MOS}, T_{60}, E_{dr} e σ_r^2, respectivamente. Comparando os resultados obtidos para a versão modificada com os

parâmetros da base não processada, percebe-se que há melhora de 2% e 34%, porém há piora de 21% e 31% para as médias de Q_{MOS}, T_{60}, σ_r^2 e E_{dr}, respectivamente.

6.5 Combinação das modificações

Foram selecionadas 20 configurações referentes a modificações no estágio de filtragem inversa, sendo as 19 configurações com os maiores valores de Q_{MOS} mais a configuração original. Foram também selecionadas 20 configurações referentes a modificações no estágio de subtração espectral, sendo as 19 configurações com os maiores valores de Q_{MOS} mais a configuração original.

Fez-se, então, a combinação das 20 configurações do estágio de filtragem inversa com as 20 configurações do estágio de subtração espectral, resultando em 400 configurações. A configuração com o melhor desempenho, obteve um valor $Q_{\mathrm{MOS}} = 3{,}71$ igual ao desempenho obtido pela melhor configuração referente à otimização somente do estágio de filtragem inversa.

As Tabelas 6.3 e 6.4 mostram o desempenho da melhor configuração referente à otimização somente do estágio de filtragem inversa (FI), da melhor configuração referente à otimização somente do estágio de subtração espectral (SE) e da melhor configuração referente à combinação das 19+1 melhores configurações de cada estágio (FI+SE), para as bases de treinamento e para os 200 sinais não anecoicos da base NBP, respectivamente. Em todos os três casos, os sinais foram submetidos aos dois estágios do algoritmo, na ordem originalmente proposta. O desempenho é representado pelo valor médio da métrica Q_{MOS}, dos parâmetros T_{60}, σ_r^2 e E_{dr}.

Tabela 6.3: Desempenho médio das medidas de avaliação de qualidade utilizando a base de treinamento

Medidas de qualidade	Base de dados não processada	Algoritmo de 2 estágios			
		Original	FI	SE	FI+SE
Q_{MOS}	3,46	3,46	3,71	3,58	3,71
T_{60} [ms]	440	354	273	247	273
σ_r^2	5,50	6,95	6,00	6,93	5,96
E_{dr}	5,54	1,97	3,06	1,81	3,06

Observando-se as Tabelas 6.3 e 6.4 nota-se que não houve melhoria adicional com a combinação das configurações referentes às otimizações individuais dos estágios. Nota-se também que em todos os casos observados o valor médio de T_{60} melhora (no caso de T_{60} quanto menor o valor, melhor é a métrica perceptual Q_{MOS}), porém tanto o valor médio de σ_r^2 quanto o valor médio de E_{dr} pioram, isto é, um aumenta e o outro diminui, respectivamente.

As Figs. 6.9 e 6.10 mostram a variação ΔT_{60}, $\Delta\sigma_r^2$ e ΔE_{dr} das estimativas de

Tabela 6.4: Desempenho médio das medidas de avaliação de qualidade utilizando os 200 sinais não anecoicos da base NBP

Medidas de qualidade	Base de dados não processada	Algoritmo de 2 estágios			
		Original	FI	SE	FI+SE
Q_{MOS}	3,36	3,42	3,52	3,43	3,52
T_{60} [ms]	517	337	368	340	373
σ_r^2	5,61	6,80	6,05	6,76	6,02
E_{dr}	7,59	2,33	4,61	2,36	4,66

T_{60}, σ_r^2 e E_{dr} utilizando as configurações original e FI+SE, respectivamente, em relação aos valores estimados para a base de dados não processada para cada um dos 18 sinais da sub-base utilizada. Essas figuras confirmam o motivo de não haver melhoria com a combinação dos dois conjuntos de propostas, uma vez que o estágio de filtragem inversa degrada os parâmetros σ_r^2 e E_{dr} de tal forma que o estágio de subtração espectral não consegue compensar de maneira efetiva essas degradações.

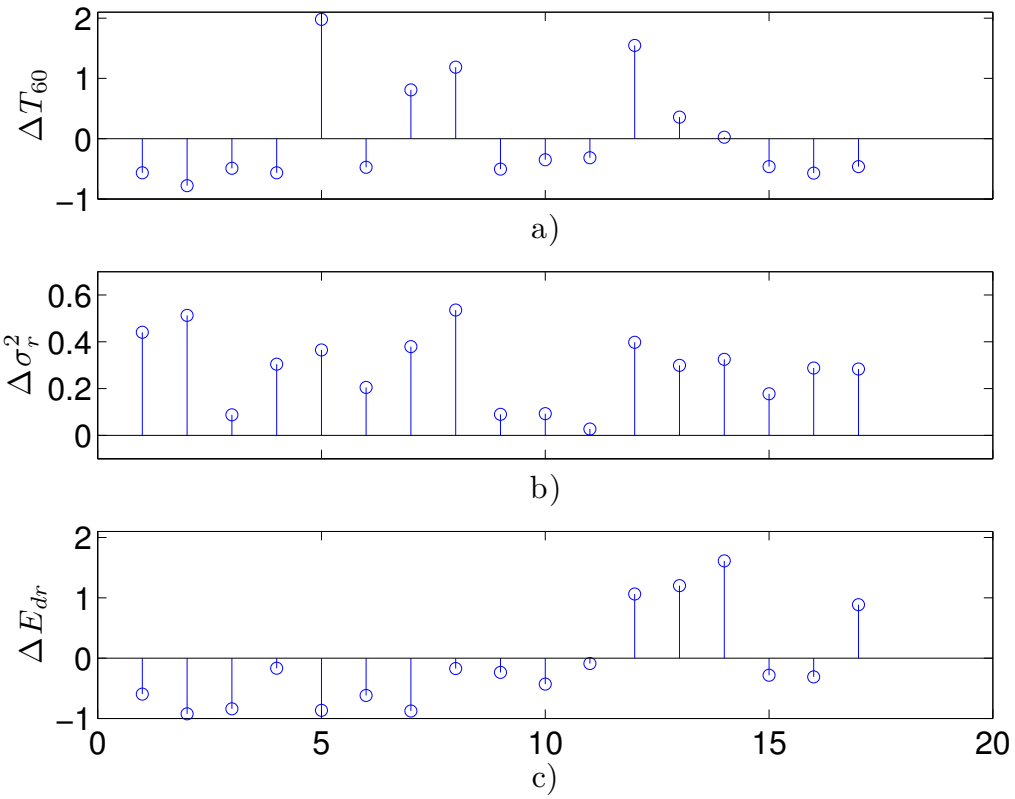


Figura 6.9: Variação relativa das estimativas de T_{60}, σ_r^2 e E_{dr} utilizando a configuração original em relação à base de dados não processada para cada um dos 18 sinais da sub-base utilizada.

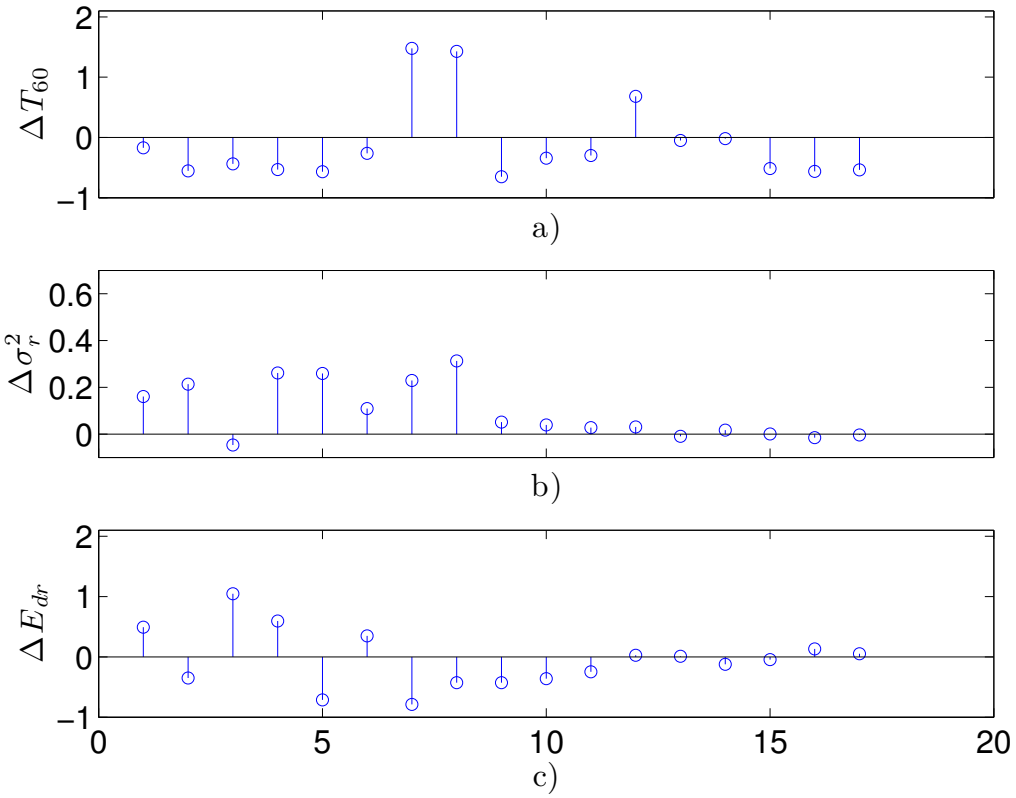


Figura 6.10: Variação relativa das estimativas de T_{60}, σ_r^2 e E_{dr} utilizando a configuração FI+SE em relação à base de dados não processada para cada um dos 18 sinais da sub-base utilizada.

6.6 Outros parâmetros para a estimação de reverberação

De acordo com Yegnanarayana e Murthy [65], e Gillespie et al. [63], os parâmetros assimetria (*skewness*), curtose e entropia podem ser utilizados para estimar a quantidade de reverberação presente em sinais de fala. Intuitivamente, quanto mais acurada for a estimação, espera-se que a filtragem inversa seja mais eficiente, como mostra [63]. Wu e Wang [3] se basearem em [63], utilizando a curtose como parâmetro a ser maximizado.

Ainda segundo Yegnanarayana e Murthy [65], especialmente em segmentos curtos, as amostras do resíduo de predição linear $s_{rp}(n)$ são menos correlacionadas do que as amostras de $s_r(n)$. Por essa razão, em particular, as métricas a seguir serão estimadas a partir de $s_{rp}(n)$.

6.6.1 Assimetria

A assimetria de uma distribuição de probabilidade de uma variável real aleatória quantifica o quão assimétrica é essa distribuição em torno da sua média. Segundo

Pääjärvi e LeBlanc [66], a assimetria do m-ésimo segmento $\chi(l)$ é definida como

$$\chi(l) = \frac{\mathbf{E}[\mathbf{s_{rp}}^3(l)]}{\mathbf{E}^2[\mathbf{s_{rp}}^2(l)]^{3/2}}. \tag{6.14}$$

6.6.2 Curtose

A curtose quantifica a altura relativa dos picos de uma distribuição de probabilidade de uma variável real aleatória. Existem diversas definições na literatura, dentre as quais duas foram escolhidas:

- A primeira definição utilizada é definida, em [63], como

$$\mathcal{K}_1(l) = \frac{\mathbf{E}[\mathbf{s_{rp}}^4(l)]}{\mathbf{E}^2[\mathbf{s_{rp}}^2(l)]} - 3. \tag{6.15}$$

  $\mathcal{K}_1(l)$ foi utilizado por Wu e Wang [3] no estágio de filtragem inversa, descrito na Seção 6.2.1, como mostrado em (6.4).

- A segunda definição considerada, dada em [67], é

$$\mathcal{K}_2(l) = \mathbf{E}[\mathbf{s_{rp}}^4(l)] - 3\mathbf{E}^2[\mathbf{s_{rp}}^2(l)]. \tag{6.16}$$

6.6.3 Entropia

A entropia é uma medida de incerteza associada a uma variável aleatória, geralmente referindo-se a entropia de Shannon [68]. De acordo com Yegnanarayana e Murthy [65], a entropia $\mathcal{H}(l)$ do l-ésimo segmento de $\mathbf{s}_{rp}(l)$ pode ser estimada a partir de um histograma com $B = 7$ *bins* das amostras de $\mathbf{s}_{rp}(l)$, gerando as probabilidades estimadas $p(i, l)$ para cada *bin* i, com $0 \leq i \leq B$, de modo que

$$\mathcal{H}(l) = -\sum_{i=1}^{B} p(i, l) \log(p(i, l)). \tag{6.17}$$

6.6.4 Análise comparativa

A capacidade de estimar a qualidade percebida de sinais de fala reverberantes das métricas anteriormente descritas foi estimada a partir do coeficiente de correlação ρ entre as estimativas dessas métricas e as notas MOS para os sinais da base NBP.

O coeficiente de correlação é mostrado na Tabela 6.5 para as métricas χ, $\mathcal{K}_1$, $\mathcal{K}_2$ e $\mathcal{H}$, representado por ρ_χ, $\rho_{\mathcal{K}_1}$, $\rho_{\mathcal{K}_2}$ e $\rho_{\mathcal{H}}$, respectivamente. Esses coeficientes de correlação foram calculados utilizando-se tamanho $R = \{10; 30; 50\}$ amostras do filtro LP, tamanho $M = \{0{,}01F_s; 0{,}032F_s\}$ amostras do segmento e sobreposição $V = \{0; 0{,}5L\}$ amostras entre segmentos adjacentes.

O melhor desempenho obtido dentre todas as métricas foi $\rho = 72\%$ (destacado em **negrito** na Tabela 6.5) relativo a métrica $\mathcal{H}$ nas configurações $R = \{30; 50\}$ amostras, $M = 0{,}032F_s$ amostras e $V = 0$ amostras. Esse resultado indica que a entropia de Shannon [68] possui a maior capacidade em estimar a qualidade percebida de sinais de fala reverberantes dentre todas as métricas testadas, sendo a mais apropriada para ser usada como função custo para o estágio de filtragem inversa.

Tabela 6.5: Coeficientes de correlação ρ_χ, $\rho_{\mathcal{K}_1}$, $\rho_{\mathcal{K}_2}$ e $\rho_\mathcal{H}$ entra a nota subjetiva MOS e as métricas χ, $\mathcal{K}_1$, $\mathcal{K}_2$, $\mathcal{H}$, respectivamente, para a base NBP

R	M	V	ρ_χ [%]	$\rho_{\mathcal{K}_1}$ [%]	$\rho_{\mathcal{K}_2}$ [%]	$\rho_\mathcal{H}$ [%]
10	$0{,}01F_s$	0	37	28	48	63
10	$0{,}01F_s$	$0{,}5L$	37	27	49	62
10	$0{,}032F_s$	0	21	19	32	69
10	$0{,}032F_s$	$0{,}5L$	22	18	31	67
30	$0{,}01F_s$	0	35	28	50	67
30	$0{,}01F_s$	$0{,}5L$	36	27	49	66
30	$0{,}032F_s$	0	20	17	35	**72**
30	$0{,}032F_s$	$0{,}5L$	21	16	35	71
50	$0{,}01F_s$	0	40	26	49	68
50	$0{,}01F_s$	$0{,}5L$	40	26	49	68
50	$0{,}032F_s$	0	25	15	35	**72**
50	$0{,}032F_s$	$0{,}5L$	27	14	35	71

6.7 Conclusões

Este capítulo propôs uma metodologia para o aperfeiçoamento de técnicas para a desreverberação de sinais de fala. Para demonstrar essa metodologia, utilizou-se o algoritmo de desreverberação de dois estágios para sinais de fala proposto por Wu e Wang [3].

Dois conjuntos de propostas de modificações no algoritmo foram analisados, gerando uma maior qualidade percebida do sinal desreverberado, sendo um conjunto referente ao estágio de filtragem inversa e o outro referente ao estágio de subtração espectral.

O primeiro conjunto de propostas consistiu na análise e escolha apropriada da ordem R do filtro de predição linear, do passo de adaptação μ do algoritmo adaptativo e na utilização de um novo critério de parada, a fim de evitar um custo computacional desnecessário. Esse conjunto de propostas aumentou em 3% a qualidade percebida de sinais de fala reverberantes estimada a partir da métrica Q_{MOS}, com uma redução de 99% aproximadamente da complexidade computacional.

O segundo conjunto de propostas consistiu na análise do fator de atenuação

ζ usado na subtração espectral, na escolha apropriada do limiar ξ, que evita o processamento de amostras com baixa energia espectral, e na otimização conjunta do tamanho da primeiras reflexões ϕ e do fator de espalhamento a da distribuição de Rayleigh. Esse conjunto de propostas aumentou em 0,3% a qualidade percebida de sinais de fala reverberantes estimada a partir da métrica Q_{MOS}.

A combinação dos dois conjuntos de propostas foi feita, porém sem melhoria na qualidade percebida estimada a partir da métrica Q_{MOS}. Isso ocorreu porque o estágio de filtragem inversa degrada os parâmetros σ_r^2 e E_{dr} de tal forma que o estágio de subtração espectral não consegue compensar de maneira efetiva essas degradações.

Para finalizar, 4 métricas foram comparadas quanto à sua capacidade de estimar a qualidade percebida de sinais de fala reverberantes, com o intuito de se obter uma indicação de qual dessas métricas seria a mais indicada para ser usada como função objetivo do estágio de filtragem adaptativa. A entropia $\mathcal{H}$ de Shannon obteve um coeficiente de correlação $\rho = 72\%$, indicando ser essa a métrica mais apropriada para tal uso.

Capítulo 7

Conclusão

7.1 Contribuições do trabalho

Este trabalho apresentou o estado da arte para a área de avaliação de qualidade de sinais reverberantes do tipo com referência, além de gerar uma base de dados composta por 204 sinais de fala reverberantes com três tipos distintos de reverberação, em que cada sinal foi avaliado por 30 sujeitos.

O Capítulo 2 definiu o problema da avaliação de qualidade percebida de sinais afetados por reverberação. Foram apresentadas as métricas subjetiva MOS e objetivas R_{DT}, de Allen, ITU-T P.563, WPESQ e SRMR, assim como uma base de dados, composta por 32 sinais reverberantes, utilizada na literatura em questão.

O Capítulo 3 contém a descrição da base de dados NBP, composta por 204 sinais de fala reverberantes, desenvolvida como parte da atual pesquisa. Este capítulo contém também a descrição detalhada do funcionamento do sistema QAreverb para a avaliação de qualidade percebida de sinais de fala reverberantes, mostrando inclusive as peculiaridades práticas do treinamento desse sistema. O desempenho do sistema QAreverb é validado com as bases NBP e MARDY, que é comparado com o desempenho das métricas apresentadas no Capítulo 2. O sistema desenvolvido mostrou-se superior a todos os demais sistemas testados, inclusive ao estado da arte (SRMR) para avaliadores sem referência.

O Capítulo 4 descreve a proposta de um algoritmo para estimar o tempo de reverberação T_{60} de um sinal reverberante de maneira cega, isto é, sem a utilização do sinal anecoico. O algoritmo descrito divide o sinal em sub-bandas a partir da DFT do sinal original e, ao final dos seus quatro estágios simples em cascata, resulta em uma estimativa do tempo de reverberação com uma correlação de mais de 92% quando empregado em duas bases de sinais de fala reverberantes.

O Capítulo 5 descreve a proposta de um método cego para a avaliação de qualidade de sinais de fala reverberantes, com base no método não cego do Capítulo 3.

Esse método se mostrou superior a todos os outros métodos cegos para estimação da qualidade percebida de sinais de fala reverberantes disponíveis na literatura e abordados neste trabalho.

O Capítulo 6 contém a descrição de dois conjuntos de propostas de modificações em um algoritmo de dois estágios para a desreverberação de sinais obtida de maneira cega, isto é, sem referência. O primeiro conjunto de propostas levou a uma melhora de 3% na qualidade estimada pela métrica proposta nesta tese se comparado ao algoritmo original nos 200 sinais não anecoicos da base NBP com uma redução da complexidade computacional de aproximadamente 99%, enquanto o segundo conjunto de propostas resultou em uma melhoria de 0,3%. Os dois conjuntos de propostas foram combinados, porém não houve melhoria na qualidade estimada, que é explicado pelo fato de o estágio de filtragem inversa degradar os parâmetros σ_r^2 e E_{dr} de tal forma que o estágio de subtração espectral não consegue compensar tais degradações. Investigou-se também qual métrica dentre assimetria, dois tipos de curtoses e a entropia de Shannon seria a mais indicada para ser usada como função objetivo do estágio de filtragem inversa. Essa investigação indicou que a melhor métrica é a entropia de Shannon, que obteve coeficiente de correção $\rho = 72\%$ com as notas MOS dos sinais da base NBP.

7.2 Próximos passos

Os seguintes processos são considerados como sequência natural do que foi concluído neste trabalho:

- Desenvolvimento de um novo estimador sem referência para a variância espectral σ_r^2.
- Utilização da entropia de Shannon como função objetivo da etapa de filtragem inversa de algoritmos de dois estágios para a desreverberação de sinais de fala.
- Uso de outras abordagens de filtragem adaptativa, além da LMS, para o desenvolvimento da etapa de filtragem inversa de algoritmos de dois estágios para a desreverberação de sinais de fala.

Publicações

Lista de artigos publicados durante o período do desenvolvimento desta tese:

1. PREGO, T. DE M., LIMA, A. A. DE, NETTO, S. L., et al. "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals", *The Journal of the Acoustical Society of America*, v. 131, n. 4, pp. 2811-2816, Apr 2012.

2. PREGO, T. DE M., LIMA, A. A. DE, NETTO, S. L. "Perceptual analysis of higher-order statistics in estimating reverberation". In: *Proc. IEEE Int. Sym. Communications, Control and Signal Processing (ISCCSP)*, pp. 1-4, Rome, Italy, Apr 2012.

3. LIMA, A. A. DE, PREGO, T. DE M., NETTO, S. L., et al. "On the quality assessment of reverberant speech", *ELSEVIER SPECOM*, v. 54, n. 3, pp. 393-401, Mar 2012.

4. PREGO, T. DE M., LIMA, A. A. DE, NETTO, S. L. "Aperfeiçoamento de algoritmo de desreverberação utilizando medidas perceptuais de qualidade". In: *Anais do Simpósio Brasileiro de Telecomunicações*, Curitiba, Brasil, Preprint 87367, Oct 2011.

5. PREGO, T. DE M., LIMA, A. A. DE, NETTO, S. L. "Perceptual improvement of a two-stage algorithm for speech dereverberation". In: *Proc. Interspeech*, pp. 209-212, Florence, Italy, Aug 2011.

6. LIMA, A. A. DE, PREGO, T. DE M., NETTO, S. L., et al. "Feature analysis for quality assessment of reverberated speech". In: *Proc. Int. Workshop Multimedia Signal Processing*, pp. 1-5, Rio de Janeiro, Brazil, Oct 2009.

Referências Bibliográficas

[1] ALLEN, J. B. "Effects of small room reverberation on subjective preference", *J. Acoust. Soc. Am.*, v. 71, n. S1, pp. S5–S5, Apr 1982.

[2] BERKLEY, D. A., ALLEN, J. B. "Acoustical Factors Affecting Hearing Aid Performance". 2nd ed., cap. Normal listening in typical rooms: the physical and psychophysical correlates of reverberation, Allyn and Bacon, 1993.

[3] WU, M., WANG, D. "A two-stage algorithm for one-microphone reverberant speech enhancement", *IEEE Trans. Audio, Speech, and Language Processing*, v. 14, n. 3, pp. 774–784, May 2006.

[4] LIMA, A. A. DE, FREELAND, F. P., ESQUEF, P. A. A., et al. "Reverberation assessment in audioband speech signals for telepresence systems". In: *Proc. Int. Conf. Signal Processing in Multimedia Applications*, pp. 257–262, Porto, Portugal, Jul 2008.

[5] MOURJOPOULOS, J., HAMMOND, J. "Modelling and enhancement of reverberant speech using an envelope convolution method". In: *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, pp. 1144–1147, Boston, USA, Apr 1983.

[6] KUTTRUFF, H. *Room Acoustics*. 4th ed. , Taylor & Francis, 2000.

[7] FIGUEIREDO, F., IAZZETTA, F. "Comparative study of measured acoustic parameters in concert halls in the city of São Paulo". In: *Proc. Int. Congress and Exposition on Noise Control Engineering*. Rio de Janeiro, Brazil, Aug 2005.

[8] KUTTRUFF, H. *Acoustics, an Introduction*. Taylor & Francis, 2007.

[9] LIMA, A. A. DE, PREGO, T. DE M., NETTO, S. L., et al. "Feature analysis for quality assessment of reverberated speech". In: *Proc. Int. Workshop Multimedia Signal Processing*, pp. 1–5, Rio de Janeiro, Brazil, Oct 2009.

[10] GOETZE, S., ALBERTIN, E., KALLINGER, M., et al. "Quality assessment for listening-room compensation algorithms". In: *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, pp. 2450–2453, Dallas, USA, 2010.

[11] STAN, G.-B., EMBRECHTS, J.-J., ARCHAMBEAU, D. "Comparison of different impulse response measurement techniques", *J. Audio Eng. Soc*, v. 50, n. 4, pp. 249–262, Apr 2002.

[12] SCHROEDER, M. R. "Integrated impulse method measuring sound decay without using impulses", *J. Acoust. Soc. Am.*, v. 65, n. 2, pp. 497–500, Aug 1979.

[13] BRIGGS, P., GODFREY, K. "Pseudorandom signals for the dynamic analysis of multivariable systems", *Proc. IEE*, v. 113, n. 7, pp. 1259–1267, Jul 1966.

[14] AOSHIMA, N. "Computer-generated pulse signal applied for sound measurement", *J. Acoust. Soc. Am.*, v. 69, n. 5, pp. 1484–1488, May 1981.

[15] FARINA, A. "Simultaneous measurement of impulse response and distortion with a swept-sine technique". In: *108^{th} AES Convention*, Paris, France, Preprint 5093, Feb 2000.

[16] RATNAM, R., JONES, D. L., WHEELER, B. C., et al. "Blind estimation of reverberation time", *J. Acoust. Soc. Am.*, v. 114, n. 5, pp. 2877–2892, Nov 2003.

[17] SABINE, W. C. *Collected Papers on Acoustics*. Harvard University Press, 1922.

[18] SCHROEDER, M. R. "New method of measuring reverberation time", *J. Acoust. Soc. Am.*, v. 37, n. 3, pp. 409–412, Mar 1965.

[19] ISO 3382. "Acoustics - measurement of the reverberation time of rooms with reference to other acoustical parameters". 2nd revision, International Organization for Standardization, 1997.

[20] CHU, W. T. "Comparison of reverberation measurements using Schroeder's impulse method and decay curve averaging method", *J. Acoust. Soc. Am.*, v. 63, n. 5, pp. 1444–1450, May 1978.

[21] XIANG, N. "Evaluation of reverberation times using a nonlinear regression approach", *J. Acoust. Soc. Am.*, v. 98, n. 4, pp. 2112–2121, Oct 1995.

[22] LUNDEBY, A., VIGRAN, T. E., BIETZ, H., et al. "Uncertainties of measurements in room acoustics", *Acustica*, v. 81, n. 4, pp. 344–355, Jul 1995.

[23] ANTSALO, P., MÄKIVIRTA, A., VÄLIMÄKI, V., et al. "Estimation of modal decay parameters from noisy response measurements". In: *Proc. Conv. Audio Engineering Society*, pp. 867–878, Amsterdam, Netherlands, May 2001.

[24] KARJALAINEN, M., ANTSALO, P., MÄKIVIRTA, A., et al. "Estimation of modal decay parameters from noisy response measurements", *J. Audio Eng. Soc.*, v. 50, n. 11, pp. 867–878, Nov 2002.

[25] JETZ, J. J. "Critical distance measurement of rooms from the sound energy spectral response", *J. Acoust. Soc. Am.*, v. 65, n. 5, pp. 1204–1211, May 1979.

[26] ZAHORIK, P. "Assessing auditory distance perception using virtual acoustics", *J. Acoust. Soc. Am.*, v. 111, n. 4, pp. 1832–1846, Apr 2002.

[27] ZAHORIK, P. "Direct-to-reverberant energy ratio sensitivity", *J. Acoust. Soc. Am.*, v. 112, n. 5, pp. 2110–2117, Nov 2002.

[28] KUSTER, M. "Reliability of estimating the room volume from a single room impulse response", *J. Acoust. Soc. Am.*, v. 124, n. 2, pp. 982–993, Aug 2008.

[29] LIMA, A. A. DE, FREELAND, F. P., JESUS, R. A. DE, et al. "On the quality assessment of sound signals". In: *Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS)*, pp. 416–419. Seattle, USA, May 2008.

[30] ZIELIŃSKI, S., RUMSEY, F. "On some biases encountered in modern audio quality listening tests - a review", *J. Audio Eng. Soc.*, v. 56, n. 6, pp. 427–451, Jun 2008.

[31] ITU-T P.800. "Methods for objective and subjective assessment of quality". International Telecommunication Union - Telecommunication Standardization Sector, 1996.

[32] WEN, J. Y. C., NAYLOR, P. A. "An evaluation measure for reverberant speech using decay tail modeling". In: *Proc. European Sig. Proc. Conf. (EUSIPCO)*, Florence, Italy, Preprint 1568981776, Sep 2006.

[33] HABETS, E. A. P. "Multi-channel speech dereverberation based on a statistical model of late reverberation". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 173–176, Philadelphia, USA, Mar 2005.

[34] SYRDAL, A. K. "Aspects of a model of the auditory representation of American English vowels", *Speech Communication*, v. 4, pp. 121–135, Aug 1985.

[35] FALK, T. H., ZHENG, C., CHAN, W.-Y. "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech", *IEEE Trans. Audio, Speech, and Language Processing*, v. 18, n. 7, pp. 1766–1774, Sep 2010.

[36] PATTERSON, R. D. "Sound of a sinusoid: spectra", *J. Acoust. Soc. Am.*, v. 96, n. 3, pp. 1409–1418, Sep 1994.

[37] ITU-T Rec. P.862.2. "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs". International Telecommunication Union - Telecommunication Standardization Sector, 2005.

[38] ITU-T P.862. "Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs". International Telecommunication Union - Telecommunication Standardization Sector, 2001.

[39] ITU-T Rec. P.563. "Single-ended method for objective speech quality assessment in narrow-band telephony applications". International Telecommunication Union - Telecommunication Standardization Sector, 2004.

[40] C.WEN, J. Y., GAUBITCH, N. D., HABETS, E. A. P., et al. "Evaluation of speech dereverberation algorithms using the MARDY database". In: *Proc. IEEE Int. Workshop Acoustic Echo and Noise Control*, Paris, France, Preprint A33, Sep 2006.

[41] FLANAGAN, J. L., JOHNSTON, J. D., ZAHN, R., et al. "Computer-steered microphone arrays for sound transduction in large rooms", *J. Acoust. Soc. Am.*, v. 78, n. 5, pp. 1508–1518, Nov 1985.

[42] ALLEN, J. B., BERKLEY, D. A. "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Am.*, v. 65, n. 4, pp. 943–950, Apr 1979.

[43] JOT, J.-M., CHAIGNE, A. "Digital delay networks for designing artificial reverberators". Preprint 3030, Feb 1991.

[44] GARDNER, W. G. "Applications of digital signal processing", chap. Reverberation Algorithms, pp. 85–131, 1998.

[45] JEUB, M., SCHÄFER, M., VARY, P. "A binaural room impulse response database for the evaluation of dereverberation algorithms". In: *Proc. 16th Int. Conf. on Digital Signal Processing*, pp. 550–554, Santorini, Greece, Jul 2009.

[46] COLE, D., MOODY, M., SRIDHARAN, S. "Intelligibility of reverberant speech enhanced by inversion of room response". In: *Proc. Int. Symp. on Speech, Image Processing, and Neural Networks*, pp. 241–244, Hong Kong, Apr 1994.

[47] GRIESINGER, D. "The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment". In: *126th AES Convention*, Munich, Germany, Preprint 7724, May 2009.

[48] LARSEN, E., IYER, N., LANSING, C. R., et al. "On the minimum audible difference in direct-to-reverberant energy ratio", *J. Acoustic. Soc. Am.*, v. 124, n. 1, pp. 450–461, Jul 2008.

[49] KAY, S. M. *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prenctice-Hall, 1993.

[50] OF THE ROYAL STATISTICAL SOCIETY, R. A. F. J. "On the interpretation of χ^2 from contingency tables, and the calculation of P", v. 85, n. 1, pp. 87–94, Jan 1922.

[51] RATNAM, R., JONES, D. L., W. D. O'BRIEN, J. "Fast algorithms for blind estimation of reverberation time", *IEEE Signal Processing Letters*, v. 11, n. 6, pp. 537–540, Jun 2004.

[52] VIEIRA, J. "Automatic estimation of reverberation time". In: *116th AES Convention*, Berlin, Germany, Preprint 6107, May 2004.

[53] VIEIRA, J. "Estimation of reverberation time without test signals". In: *118th AES Convention*, Barcelona, Spain, Preprint 6499, May 2005.

[54] WEN, J. Y. C., HABETS, E. A. P., NAYLOR, P. A. "Blind estimation of reverberation time based on the distribution of signal decay rates". In: *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 329–332, Las Vegas, USA, Mar 2008.

[55] LÖLLMANN, H. W., VARY, P. "Estimation of the reverberation time in noisy environments". In: *Proc. IEEE Int. Workshop Acoustic Echo and Noise Control*, Seattle, USA, Preprint 9033, Sep 2008.

[56] FALK, T. H., CHAN, W.-Y. "A non-intrusive quality measure of dereverberated speech". In: *Proc. IEEE Int. Workshop Acoustic Echo and Noise Control*, Seattle, USA, Preprint 9009, Sep 2008.

[57] BERANEK, L. L. "Concert hall acoustics - 1992", *J. Acoustic. Soc. Am.*, v. 92, n. 1, pp. 1–39, Jul 1992.

[58] HABETS, E. A. P., GANNOT, S., COHEN, I. "Late reverberant spectral variance estimation based on a statistical model", *IEEE Signal Processing Letters*, v. 16, n. 9, pp. 770–773, Sep 2009.

[59] KIEFNER, J. "Sequential minimax search for a maximum". In: *Proc. Amer. Math. Soc.*, v. 4, pp. 502–506, Jun 1953.

[60] DAU, T., PUSCHEL, D., KOHLRAUSCH, A. "A quantitative model of the effective signal processing in auditory system. I - model structure", *J. Acoustic. Soc. Am.*, v. 99, pp. 3615–3622, Jun 1996.

[61] APPEL, R., BEERENDS, J. "On the quality of hearing one's own voice", *J. Audio Eng. Soc.*, v. 50, n. 4, pp. 237–248, Apr 2002.

[62] MOORER, J. A. "About this reverberation business", *Computer Music Journal*, v. 3, n. 2, pp. 13–28, Jun 1979.

[63] GILLESPIE, B. W., MALVAR, H. S., FLORÊNCIO, D. A. F. "Speech dereverberation via maximum-kurtosis subband adaptive filtering". In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3701–3704, Salt Lake, USA, May 2001.

[64] HAYKIN, S. *Adaptive Filters Theory*. 4th ed. Upper Saddle River,N.J., Prentice-Hall, 2002.

[65] YEGNANARAYANA, B., MURTHY, P. S. "Enhancement of reverberant speech using LP residual signal", *IEEE Trans. on Speech and Audio Proc.*, v. 8, n. 3, pp. 267–281, May 2000.

[66] PÄÄJÄRVI, P., LEBLANC, J. P. "Skewness maximization for impulsive sources in blind deconvolution". In: *Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG)*, pp. 304–307, Espoo, Finland, Apr 2004.

[67] TANRIKULU, O., CONSTANTINIDES, A. G. "The LMK algorithm with time-varying forgetting factor for adaptive system identification in additive output-noise". In: *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, pp. 1850–1853, Atlanta, USA, May 1996.

[68] SHANNON, C. E. "A mathematical theory of communication", *Bell System Technical Journal*, v. 27, n. 3, pp. 379–423, Jul 1948.

[69] LIMA, A. A. DE, NETTO, S. L., BISCAINHO, L. W. P., et al. "Quality evaluation of reverberation in audioband speech signals". In: Filipe, J., Obaidat, M. S. (Eds.), *e-Business and Telecommunications*, v. 48, *Communications in Computer and Information Science*, Springer Berlin Heidelberg, pp. 384–396, 2009. ISBN: 978-3-642-05197-5.

[70] WU, M., WANG, D. "A pitch-based method for the estimation of short reverberation time", *Acta Acustica united with Acustica*, v. 92, n. 2, pp. 337–339, Mar 2006.

[71] BISPO, B. C., ESQUEF, P. A. A., BISCAINHO, L. W. P., et al. "EW-PESQ: a quality assessment method for speech signals sampled at 48 kHz", *J. Audio Eng. Soc*, v. 58, n. 4, pp. 251–268, Apr 2010.

Apêndice A

Salas utilizadas nas gravações da NBP

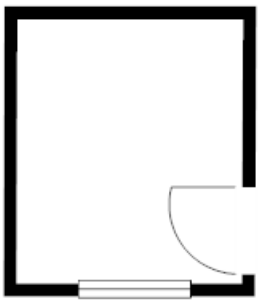


Figura A.1: Desenho esquemático da sala cabine do efeito de reverberação real. Dimensões (C, L, A) em metros 2,1×1,8×2,4.

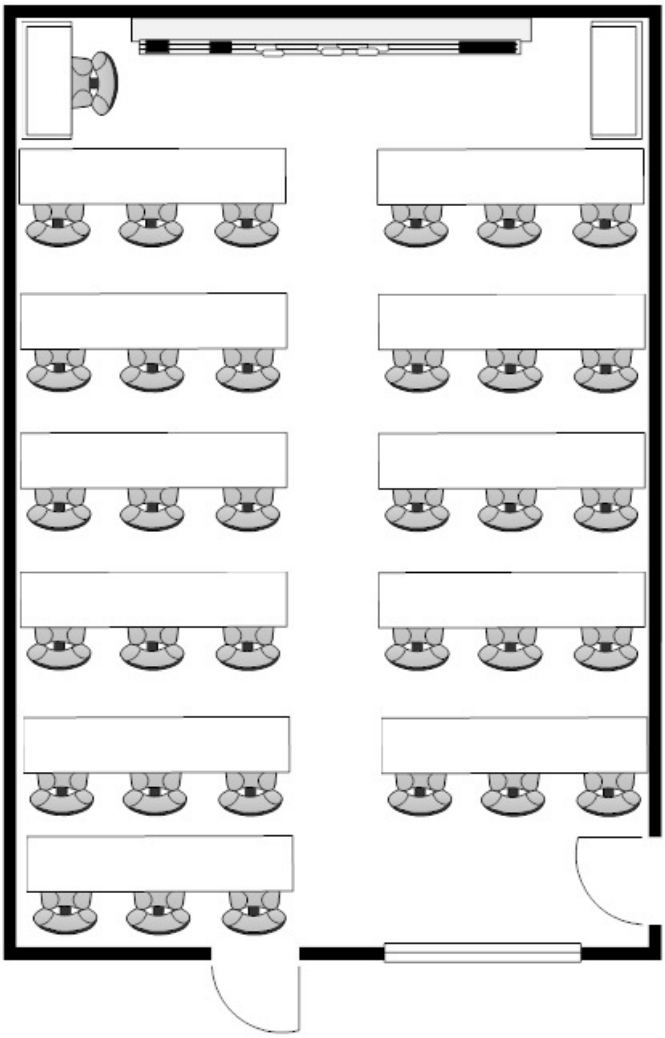


Figura A.2: Desenho esquemático da sala escritório1 do efeito de reverberação real. Dimensões (C, L, A) em metros 7,4×5,0×2,7.

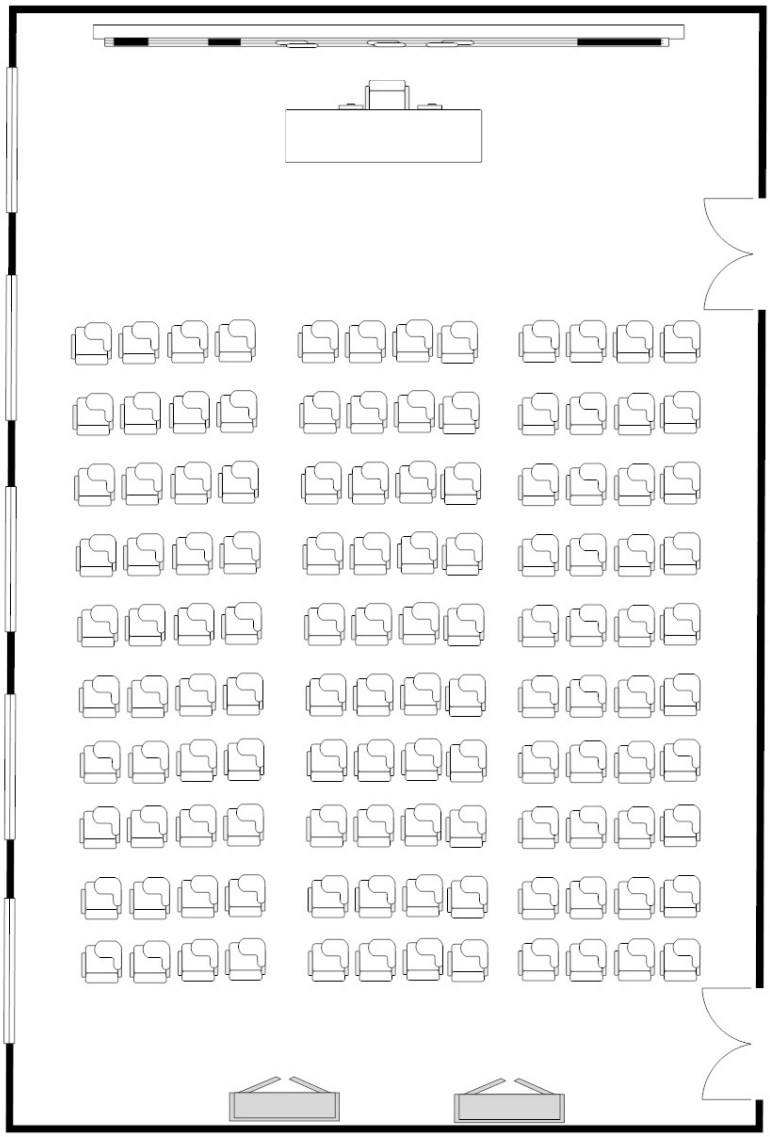


Figura A.3: Desenho esquemático da sala aula1 do efeito de reverberação real. Dimensões (C, L, A) em metros 15,0×10,0×4,0.

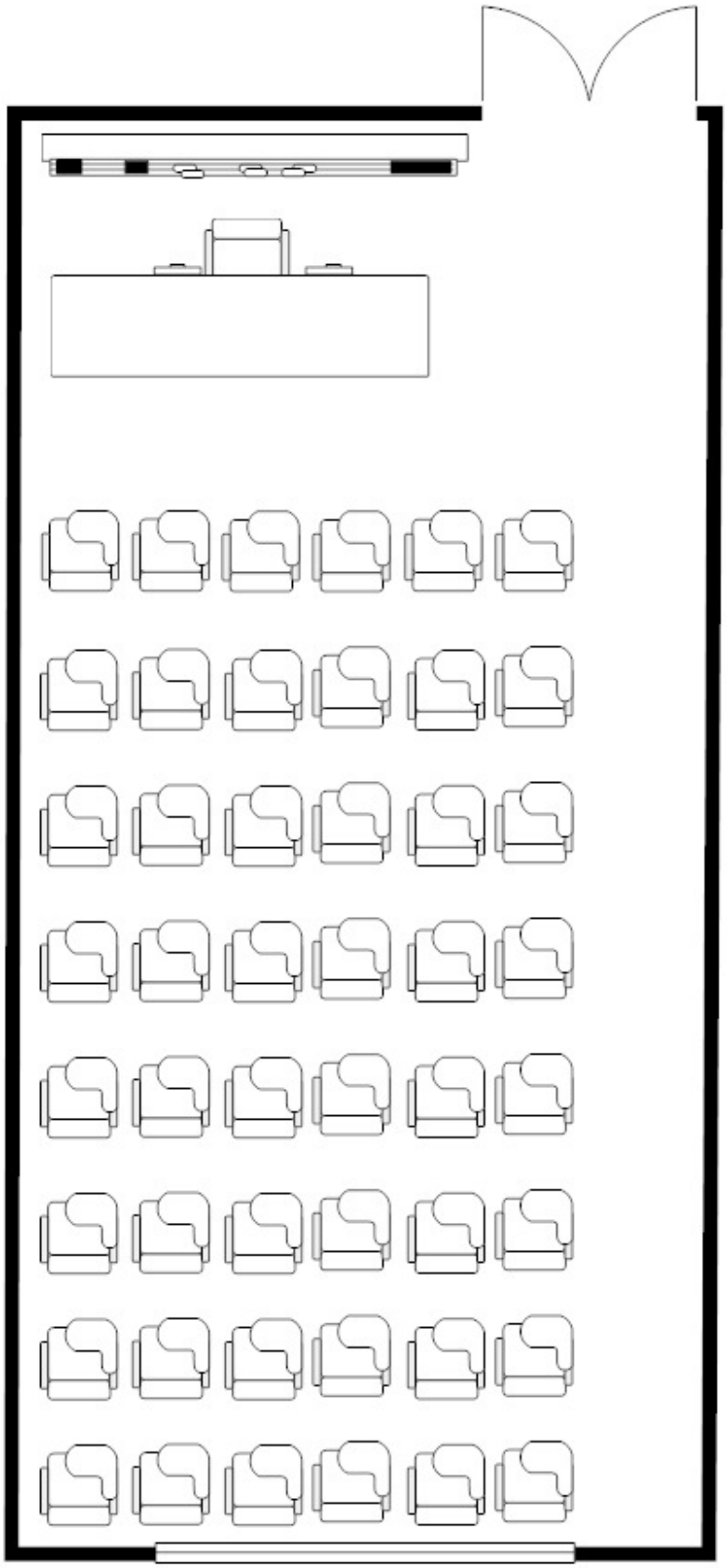


Figura A.4: Desenho esquemático da sala reunião1 do efeito de reverberação real. Dimensões (C, L, A) em metros 10,0×4,8×3,2.

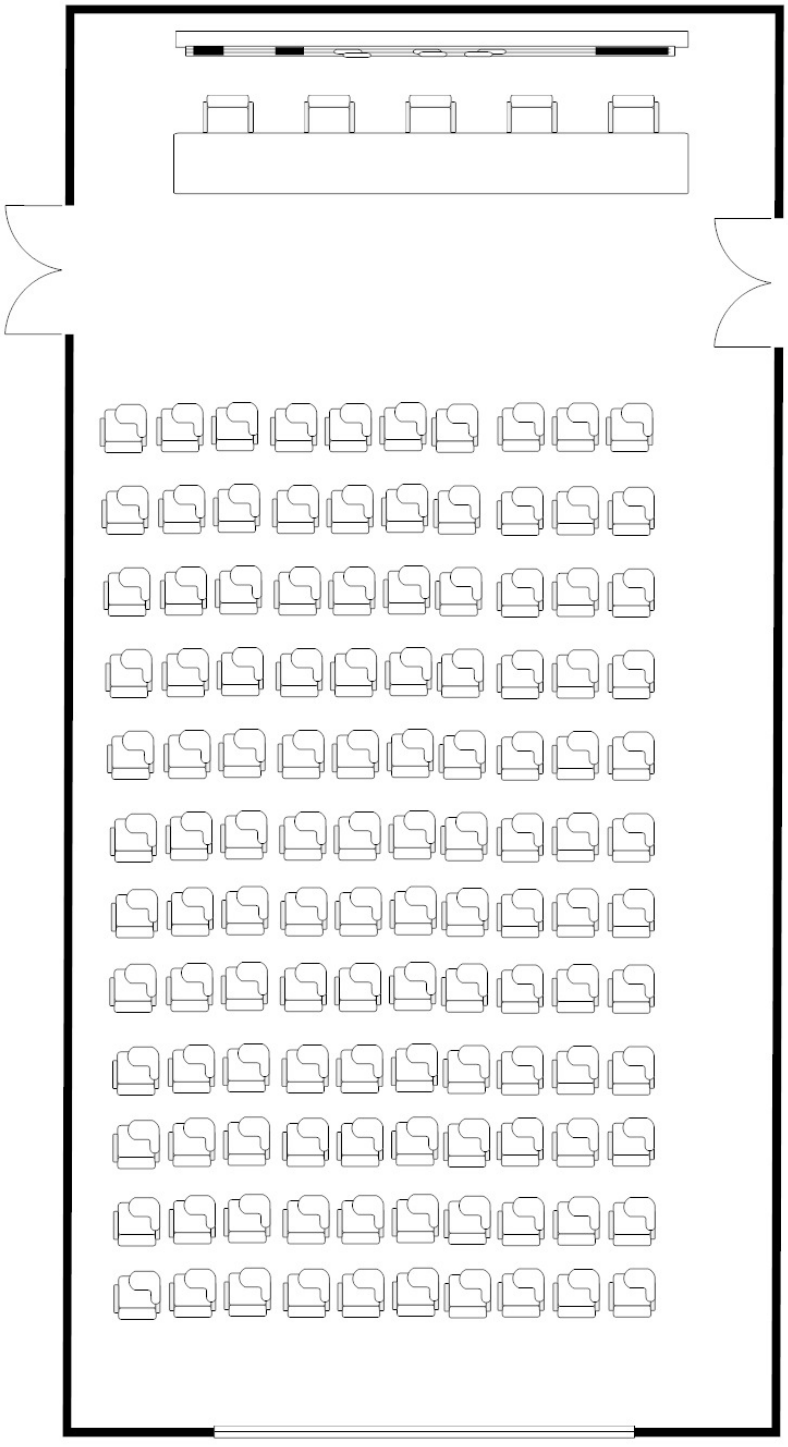


Figura A.5: Desenho esquemático da sala aula2 do efeito de reverberação real. Dimensões (C, L, A) em metros 16,5×8,2×3,5.

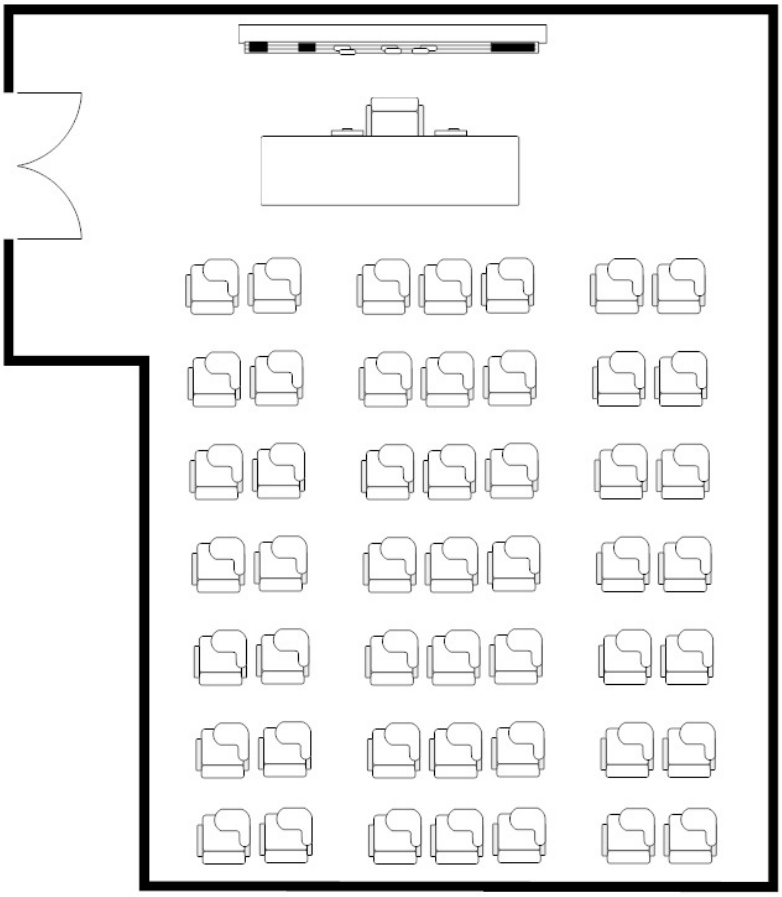


Figura A.6: Desenho esquemático da sala reunião2 do efeito de reverberação real. Dimensões (C, L, A) em metros 9,0×7,3×3,5.

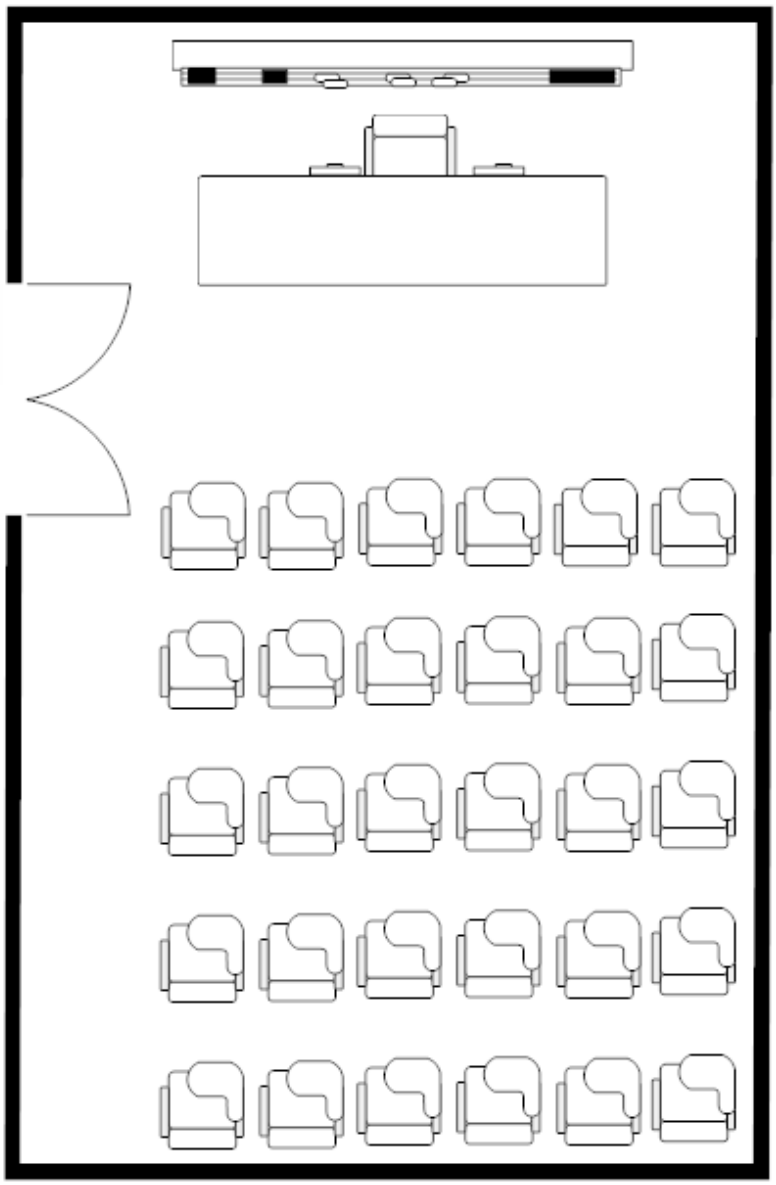
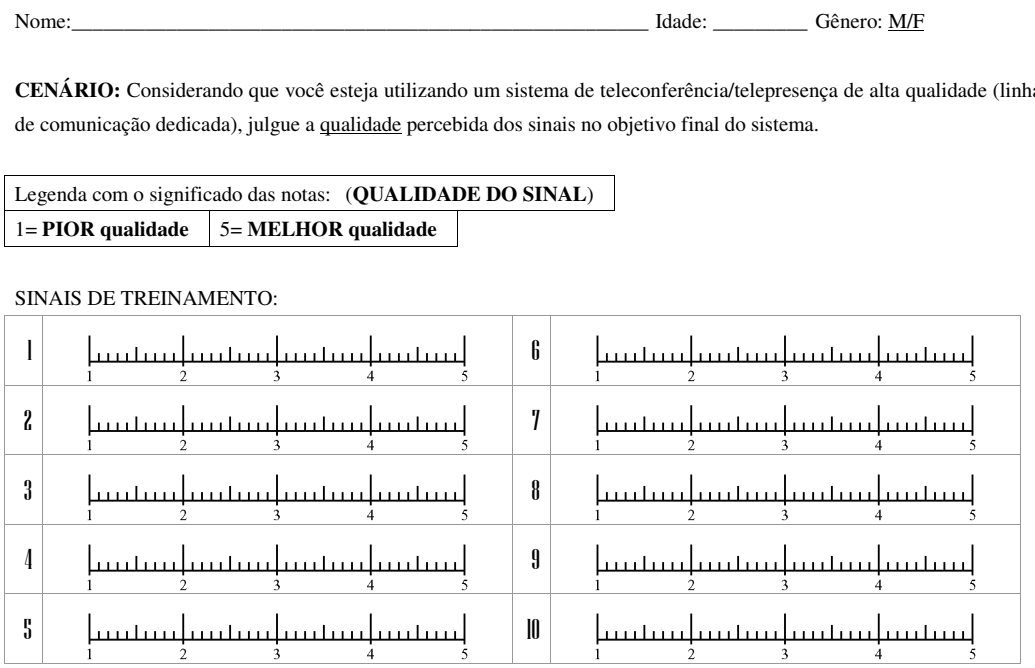


Figura A.7: Desenho esquemático da sala escritório2 do efeito de reverberação real. Dimensões (C, L, A) em metros 7,4×4,8×4,3.

Apêndice B

Formulário do teste subjetivo da base NBP



Nome:_____ Idade: _____ Gênero: M/F

CENÁRIO: Considerando que você esteja utilizando um sistema de teleconferência/telepresença de alta qualidade (linha de comunicação dedicada), julgue a qualidade percebida dos sinais no objetivo final do sistema.

Legenda com o significado das notas: (**QUALIDADE DO SINAL**)	
1= **PIOR qualidade**	5= **MELHOR qualidade**

SINAIS DE TREINAMENTO:

1		6	
2		7	
3		8	
4		9	
5		10	

Figura B.1: Página 1 do formulário do teste subjetivo para os sinais da base NBP.